CONSISTENT BI-LEVEL VARIABLE SELECTION VIA COMPOSITE GROUP BRIDGE PENALIZED REGRESSION

by

INDU SEETHARAMAN

B.S., University of Madras, India, 2001
M.B.A., University of Madras, India, 2003

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2013

Approved by:

Major Professor
Dr. Kun Chen

Copyright

Abstract

We study the composite group bridge penalized regression methods for conducting bi-level variable selection in high dimensional linear regression models with a diverging number of predictors. The proposed method combines the ideas of bridge regression (Huang *et al.*, 2008a) and group bridge regression (Huang *et al.*, 2009), to achieve variable selection consistency in both individual and group levels simultaneously, i.e., the important groups and the important individual variables within each group can both be correctly identified with probability approaching to one as the sample size increases to infinity. The method takes full advantage of the prior grouping information, and the established bi-level oracle properties ensure that the method is immune to possible group misidentification. A related adaptive group bridge estimator, which uses adaptive penalization for improving bi-level selection, is also investigated. Simulation studies show that the proposed methods have superior performance in comparison to many existing methods.

Key Words and Phrases: Bi-level variable selection; High-dimensional data; Oracle property; Penalized regression; Sparse models.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My sincere gratitude to my major professor, Dr. Kun Chen, for his constant motivation, guidance, support and encouragement. His passion towards the discipline of statistics is deeply inspiring and infectious. His careful review of several versions of this manuscript improved the quality of this report. Without him, this work would not have been possible.

I would like to extend my thanks to Dr. Gary Gadbury and Dr. Weixin Yao for their willingness to serve on my committee and for their valuable time and insightful comments.

I would like to dedicate this report to my family for their love, encouragement and prayers. My grandmother inspires me with her patience and kindness, my parents with their hardwork, honesty and moral strength and, my sister with her encouragement and insights. Last but not the least, my loving gratitude to my husband, who with his calming presence, ever loving support and a very nice sense of humor makes my dreams come true.

Chapter 1

Introduction

In contemporary scientific research, high dimensional data has become increasingly common in various fields including genetics, finance, medical imaging, social networks, etc. Complex statistical models of high dimensionality, e.g., regression models with large number of predictors, are routinely formulated. One key of high dimensional modeling is to efficiently conduct dimension reduction so that a parsimonious model can be built with both strong predictive power and clear interpretation. In recent years, many penalized estimation methods have been proposed, which are capable of conducting efficient variable selection and model estimation simultaneously. To list a few, Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), SCAD (Fan and Li, 2001), bridge regression and MCP (Zhang, 2010) were designed for individual-level variable selection. Group Lasso (Yuan and Lin, 2006), group MCP (Zhang, 2007) and group SCAD (Wang *et al.*, 2007) were proposed for group-level variable selection in the presence of some prior grouping structure among variables. For a comprehensive account of the developments of variable selection techniques, see, e.g., Buhlmann and van de Geer (2009) and Huang *et al.* (2012).

In many applications, however, it is desirable to conduct both group-level and individual-level variable selection, i.e., not only we want to identify which groups of variables contain useful information, but also we want to identify the truly important variables within each selected group. For example, an impact study was designed to determine the effects of different risk factors on body mass index (BMI) of high school students in two Seattle

public schools (Huang *et al.*, 2009). The predictors in the study could be naturally divided into several groups based on the types of the measurements. Particularly, a set of dummy variables are created to represent ethnicity; it is thus of interest to know whether ethnicity is an important risk factor on BMI, and if so, which ethnicity groups show significantly different effects. In genetic studies, gene expression profiles are commonly served as high dimensional covariates to predict cancer risk. Often the genes can be grouped based on pathways; therefore, it is of importance to be able to both select relevant pathways and identify the useful genes within each selected pathway. To some extent, the need for bi-level selection is also motivated by the fact that in practice, we rarely possess exactly correct group information such that the important variables and unimportant ones are entirely separated to form different groups. Therefore, a variable selection method should allow flexible incorporation of the prior group information, so that not only can the method take full advantage of the grouping structure when it is correct, but also the method can be immune from possible group misspecification.

Motivated by these application needs, Huang *et al.* (2009) propose the group bridge regression approach. The method penalizes the L_1 norms of each group of coefficients using a bridge penalty, to induce sparsity among the regression coefficients at both the group level and the individual level. Huang *et al.* (2009) showed that under suitable conditions, the method enjoys oracle properties at group selection. However, it is well known that using an L_1 norm penalty often leads to overselection, unless strong and maybe unrealistic assumptions are imposed (Buhlmann and van de Geer, 2009). As a consequence, a disadvantage of the group bridge method is that it does not perform well at individual level variable selection, exhibiting similar behaviors as Lasso.

Combining the ideas of bridge regression, adaptive Lasso and group bridge method, we propose and study the composite group bridge (CoGB) and the adaptive group bridge (AdGB) methods for bi-level selection. Unlike the group bridge method in which a L_1 penalty is used to induce within-group sparsity, we adopt either a nonconvex bridge penalty

or an adaptively weighted L_1 penalty. We show that under suitable conditions, the proposed methods achieve the oracle properties for both group and individual level variable selection, allowing a diverging number of predictors. To our knowledge, this is the first time that bi-level selection consistency is rigorously developed.

The rest of the report is organized as follows. In Chapter 2, we discuss the need for dimension reduction in high dimensional data analysis and provide a brief review of the existing techniques for shrinkage estimation and variable selection. We propose the CoGB method and develop an efficient computation algorithm for its optimization in Chapter 3. We also propose the simpler AdGB method, when some reliable initial estimator is available for constructing the adaptive weights. Asymptotic properties of the CoGB estimator is studied in Chapter 4. Simulation studies illustrating the new methods are presented in Chapter 5.

Chapter 2

Literature Review

2.1 Variable Selection in Multiple Linear Regression

Consider the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} = \mathbf{x}_1\beta_{01} + \cdots + \mathbf{x}_d\beta_{0d} + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

where $\mathbf{y} = (y_1, ..., y_n)'$ is the response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ the design matrix, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d})'$ a vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ an error vector consisting of independently and identically distributed (i.i.d.) random errors with mean 0 and variance σ_ε^2 . Without loss of generality, we assume that the response is centered, i.e., $\sum_{i=1}^n y_i = 0$, and the predictors are standardized, i.e., $\sum_{i=1}^n x_{ik} = 0$ and $\sum_{i=1}^n x_{ik}^2 = n$, for $1 \leq k \leq d$. So there is no intercept term in the model.

In many applications, the number of predictors d may be comparable to or even exceed the sample size. The commonly used least squares estimator of $\boldsymbol{\beta}_0$, which is obtained by minimizing the residual sum of squared error (RSS),

$$RSS(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k\beta_k\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.1.2}$$

often leads to overfitting and poor predictive performance. Here and henceforth, we use $\|\cdot\|_q$ to denote the L_q norm, for $q > 0$. Moreover, it can be difficult to interpret the fitted model from the least squares estimation when the number of predictors is large. To improve predictive accuracy and facilitate model interpretation, it is often imperative to conduct

variable selection, so that a parsimonious regression model with only a few selected leading predictors can be built.

A natural idea of producing a good model with a reduced number of predictors is best subset selection, which aims to exhaustively search over all 2^d possible models and select the best one judged by some information criterion such as AIC (Akaike, 1974), Mallows C_P (Mallows, 1973), BIC (Schwarz, 1978), etc. Although the idea is appealing, this procedure is infeasible to implement when d is large, as its computational complexity grows exponentially with d. Furthermore, the method suffers from instability due to sampling variability and discontinuity (Breiman, 1996). To get around the computation issue, some commonly used model selection procedures in practice include forward selection, backward elimination and stepwise selection. The basis ideas are as follows.

- Backward elimination begins with the full set of variables and at each step, sequentially drops a variable that is deemed to be the least important based on some information criterion or some hypothesis testing procedure. This process is repeated until all variables having nonsignificant effects are dropped from the model.

- Forward selection starts with an empty set of variables and sequentially adds a variable to the model in each step. Again, the choice of the most important variable in each step can be based on some information criterion or some hypothesis testing procedure. This procedure is repeated until no new predictors can be added to the model.

- A hybrid stepwise selection procedure also starts with the null model of no predictor. At each step, after adding a variable, it then tries to eliminate any variable that is determined to be insignificant in the current model. The alternation between the selection and elimination steps is continued until all variables have either been retained for inclusion or removed. (Izenman, 2008).

A common criticism of the above three procedures is that only a small subset of possible models is visited and compared during the model-building process, so the resulting model

may be suboptimal. Nevertheless, the discrete nature of these procedures makes them suffer from instability.

In recent years, the celebrated penalized estimation approaches, being capable of conducting efficient and simultaneous dimension reduction and model estimation, have undergone exciting developments. The main idea of this broad class of approaches is to mitigate the curse of dimensionality by assuming that the true coefficient vector $\boldsymbol{\beta}_0$ has some low-dimensional structure, e.g., sparsity, and employing proper regularization approaches for model estimation. For Gaussian data, it is appropriate to estimate $\boldsymbol{\beta}_0$ by minimizing the penalized least squares criterion with respect to $\boldsymbol{\beta}$,

$$RSS(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) + P_\lambda(\boldsymbol{\beta}),$$

where $RSS(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$ is defined in (2.1.2), $P_\lambda(\cdot)$ some penalty function measuring the complexity of the enclosed coefficient vector, and λ a non-negative tuning parameter controlling the degree of penalization. In what follows, we shall review some commonly used and inspiring penalized estimation methods developed in this exciting era of big data.

2.2 Ridge Regression, Lasso and Related Methods

Shrinkage estimation can be performed by using ridge regression, in which the penalty function is proportional to the squared L_2 norm of the coefficient vector, i.e., the ridge estimator $\hat{\boldsymbol{\beta}}_{ridge}$ is obtained by minimizing

$$L_{Ridge}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2. \tag{2.2.1}$$

It can be shown that,

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y}.$$

Comparing to the least squares estimation, the ridge estimator is unique for any $\lambda > 0$, even when the design matrix is not of full rank. Applying ridge penalty has the effect of shrinking the estimates toward zero, which introducing bias but reducing the variance of the estimator,

and thus ridge regression is especially beneficial in the presence of multicollinearity. However, as exact sparsity is not induced among the estimated regression coefficients, ridge regression does not perform variable selection. Consequently, although using ridge estimation can alleviate overfitting and achieve better predictive performance, it may prove difficult to interpret the resulting model with a large number of predictors.

Lasso (Tibshirani, 1996) stands for "Least Absolute Shrinkage and Selection Operator". The Lasso estimator, denoted by $\hat{\boldsymbol{\beta}}_{lasso}$, is obtained by minimizing

$$L_{Lasso}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{2.2.2}$$

Comparing to ridge regression, the only change here is that an L_1 norm penalty is used instead of a squared L_2 norm penalty. Note that both ridge and Lasso methods can be cast as constrained estimation problems, i.e., $\hat{\boldsymbol{\beta}}_{ridge}$ is obtained by minimizing $RSS(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_2^2 \leq t$, and $\hat{\beta}_{lasso}$ is obtained by minimizing $RSS(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_1 \leq t$, where t is a non-negative constant corresponding to some tuning parameter λ. The difference in their penalty terms leads to important consequences. To illustrate, consider a simple case when there are only two predictors in the model, i.e., $d = 2$. Figure 2.1 shows the geometrical representations of Lasso and ridge regression methods. The constrained region for Lasso is a rotated square, while the region becomes a disk in ridge regression. The least squares estimate $\hat{\boldsymbol{\beta}}_{ls}$ is shown in both panels, surrounded by the elliptical contours of the sum of squared error $RSS(\boldsymbol{\beta})$. For the Lasso method, there is a positive probability that the contour may touch the rotated square at its corners so that a sparse estimator is produced, i.e., some coefficient can be estimated as exact zeros. On the other hand, the constrained area due to ridge penalization is smooth, and hence producing a sparse solution is of probability 0.

Lasso can be efficiently solved by several methods, e.g., a modified least angle regression algorithm (LARS) (Efron *et al.*, 2004; Park and Hastie, 2007) and the coordinate descent algorithm (CDA) (Friedman *et al.*, 2007). The theoretical properties of Lasso have been throughly investigated (Knight and Fu, 2000; Zhao and Yu, 2006; Zhang and Huang, 2008).

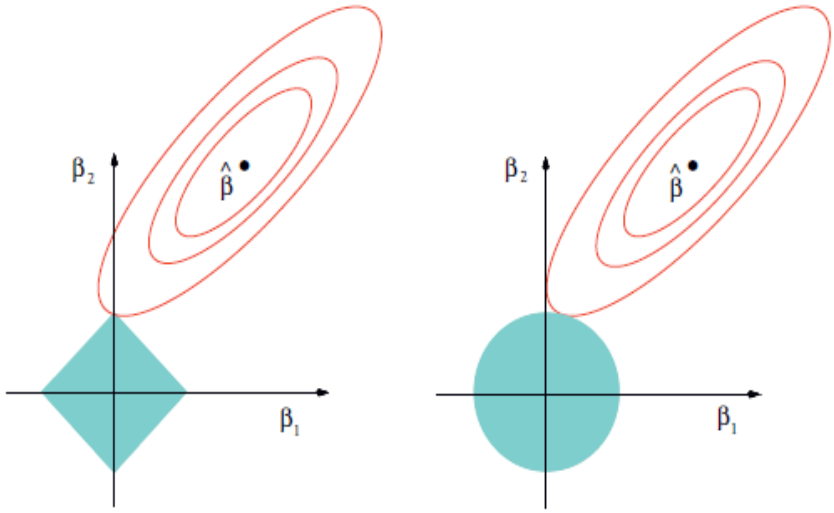


Figure 2.1: *Geometry of Lasso versus Ridge estimation for a two parameters case (Hastie et al., 2009). The solid rotated square region is the constraint region for* $|\beta_1| + |\beta_2| \leq t$ *and the solid circle region is the constraint region for* $\beta_1^2 + \beta_2^2 \leq t^2$*, where* t *is a constant. The ellipses are the contours of the least squares error function.*

Particularly, Zhao and Yu (2006) showed that under a strong irrepresentable condition, Lasso achieves the oracle properties (Fan and Li, 2001), i.e., the method identifies the correct subset of important predictors with probability tending to one as the sample size increases, and that the asymptotic distribution of the vector of nonzero coefficients is the same as it would have if the sparse model structure is known a priori. The irrepresentable condition essentially requires that the important predictors and the irrelevent ones can not be strongly related, so the irrelevent predictors that are not in the true model are in a certain sense "irrepresentable" by the predictors that are in the true model. In practice, however, it is hard to verify this condition, and empirically Lasso tends to select more predictors than needed and the estimator may exhibit large bias.

Adaptive Lasso estimator proposed by Zou (2006) is obtained by minimizing

$$L_{AdLasso}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{1,w} \tag{2.2.3}$$

where $\mathbf{w} = (w_1, \ldots, w_d)'$ is a set of adaptive weights, and we used the operator $\|\cdot\|_{1q,w}$ to

denote a $\mathbf{w}$-weighted L_1 norm, i.e., $\|\boldsymbol{\beta}\|_{1,w} = \sum_{k=1}^{d} w_k|\beta_k|$. Adaptive Lasso, as an extension of Lasso, remains to be a convex optimization problem, and it can also be solved efficiently using LARS or CDA. To construct proper weights, a natural way is to set $w_k = (\hat{\beta}_k^{(0)})^{-\gamma}$, where $\hat{\beta}_k^{(0)}$ is some initial estimator of β_{0k}, e.g., the least squares estimator or the Lasso estimator, and γ is a prespecified positive constant. As such, larger coefficients receive a lighter penalty to help reduce the bias and smaller coefficients receive a heavier penalty to improve sparsity. With the help of some well-behaved adaptive weights, Zou (2006) showed that adaptive Lasso estimator enjoys oracle properties in the sense of Fan and Li (2001), without the requirement of the irrepresentable condition, and Huang *et al.* (2008b) later extended the results for adaptive Lasso to high-dimensional models with a diverging number of predictors. Zou and Hastie (2005) and Zou and Zhang (2009) proposed the elastic net and the adaptive elastic net methods to combine the strength of L_1 and L_2 penalization.

In the penalized regression framework, the construction of the penalty function and the statistical properties of the resulting estimators have been extensively studied. The aforementioned methods all used convex penalties to promote sparsity. However, the weakness of convex penalties is well understood (Buhlmann and van de Geer, 2009). A promising way is to adopt nonconvex penalties in penalized regression, which may lead to superior properties in both model estimation and variable selection under milder conditions. Some popular choices include SCAD (Fan and Li, 2001), bridge regression (Huang *et al.*, 2008a), MCP (Zhang, 2010), etc. We refer the interested reader to Buhlmann and van de Geer (2009) for a comprehensive account of these techniques.

2.3 Group and Bi-level Selection

In many statistical modeling problems, the variables exhibit some natural grouping structures and these is a need to select them at the group level. For example, a categorical variable with multiple levels can be coded as a group of dummy variables. Using basis expansion, a nonparametric component in a regression model can be written as a linear

combination of a set of basis functions. In gene expression data analysis, gene expression profiles can be grouped according to the pathways these genes belong to. In CT-Scanned lung image studies, the lung airway measurements can be grouped by either their generation number or the type of measurements. In all these cases, it is often desirable that a group of variables are either kept or eliminated from the model together.

Several penalized estimation methods have been developed for group selection. Yuan and Lin (2006) proposed the group Lasso method as a natural extension of Lasso (Tibshirani, 1996), in which the L_2 norms of the groups of coefficients are penalized using an L_1 penalty. The group Lars and group non-negative garrote methods were also studied by Yuan and Lin (2006). Meier *et al.* (2008) studied the group Lasso in logistic regression. Zhao *et al.* (2009) proposed a composite absolute penalty, which combines the properties of norm penalties at the across-group and within-group levels to facilitate hierarchical variable selection. Other methods proposed for group level variable selection include group SCAD (Wang *et al.*, 2007), group MCP (Breheny, 2009), etc. To further facilitate model interpretation, an intelligent idea is to conduct both group-level and individual-level variable selection, i.e., not only we want to identify which groups are important, but also we want to identify the important variables within each selected group. The group bridge method proposed by Huang *et al.* (2009) was the first method developed for bi-level variable selection. Breheny (2009) proposed a general composite penalty form for bi-level selection and developed a coordinate descent algorithm for solving these problems. We refer the interested reader to Huang *et al.* (2012) for a review of the group selection techniques.

In what follows, we present the main ideas of the group Lasso and the group bridge methods to illustrate the general methodology of group section and bi-level section, respectively. Consider model (2.1.1). Let $A_1, ..., A_J$ be subsets of $\{1, ..., d\}$ representing known groupings of the predictors, and $\boldsymbol{\beta}_{A_j} = (\beta_k, k \in A_j)'$ be the vector of regression coefficients in the jth group. We assume only the first J_1 groups contain useful predictors, i.e. $\boldsymbol{\beta}_{0A_j} \neq \mathbf{0}$ for $j = 1, ..., J_1$ and $\boldsymbol{\beta}_{0A_j} = \mathbf{0}$ for $j = J_1 + 1, ..., J$.

The group Lasso method (Yuan and Lin, 2006) minimizes the following objective function,

$$L_{grLasso}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_2,$$

where c_j accounts for the varying group sizes, commonly chosen as $\sqrt{|A_j|}$, the square root of the number of predictors in group j. In the above criterion, the L_2 norm of each group of coefficients is penalized. If all the group sizes equal to one, group Lasso reduces to a Lasso optimization problem. In general case, the group Lasso penalty induces sparsity at the group level due to the L_1 norm penalization at the group level. However, similar to Lasso, the group Lasso in general can not achieve selection consistency and tends to select more groups than needed.

The group bridge approach for variable selection was introduced by Huang *et al.* (2009), in which $\boldsymbol{\beta}_0$ is estimated by minimizing

$$L_{GB}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma}, \tag{2.3.1}$$

where $\gamma \in (0, 1]$ is the bridge index, $\lambda > 0$ is the regularization parameter, and c_js are constants adjusting for group sizes, usually set as $c_j = |A_j|^{1-\gamma}$. Group bridge criterion reduces to a standard bridge criterion when all the group sizes are equal to one, and further reduces to Lasso when $\gamma = 1$. The method penalizes the L_1 norms of the groups of coefficients using a bridge penalty, to induce sparsity among the regression coefficients at both the group level and the individual level. Huang *et al.* (2009) showed that under suitable conditions, the method enjoys oracle properties at group level selection. However, it is well known that using an L_1 norm penalty often leads to overselection, unless strong assumptions are imposed on the design matrix (Buhlmann and van de Geer, 2009). As a consequence, a disadvantage of the group bridge method is that it does not perform well at individual level variable section, similar to the behavior of Lasso.

Chapter 3

Composite Group Bridge Regression

3.1 Composite Group Bridge Criterion

Consider the multiple linear regression model in (2.1.1). Recall that we assume the predictors form J groups, and only the first J_1 groups are relevant, i.e., $\boldsymbol{\beta}_{0A_j} \neq \mathbf{0}$ for $j = 1, ..., J_1$ and $\boldsymbol{\beta}_{0A_j} = \mathbf{0}$ for $j = J_1+1, ..., J$. We further assume in each of the first J_1 groups, only a subset of the predictors is important. For each A_j, $j = 1, ..., J_1$, let $A_j^1 = \{k; \beta_{0k} \neq 0, k \in A_j\}$ and $A_j^2 = \{k; \beta_{0k} = 0, k \in A_j\}$. Note that the J groups may overlap with each other and their union is allowed to be a proper subset of all the predictors.

Motivated by both bridge and group bridge penalized regression methods (Huang *et al.*, 2009), we propose to conduct bi-level variable selection and model estimation by minimizing

$$L_n(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j (\sum_{k \in A_j} |\beta_k|^{\mu})^{\gamma}, \tag{3.1.1}$$

where $\mu \in (0, 1]$, $\gamma \in (0, 1]$, c_js are group level weights adjusting for the dimensions or magnitudes of each group of coefficients, and λ_n is a tuning parameter controlling degrees of penalization. Unless otherwise noted, we set $c_j = |A_j|^{1-\gamma}$.

The proposed method subsumes and extends both the bridge and group bridge approaches. In the group bridge method, a bridge penalty is used to penalize the L_1 norm of each group of coefficients, which in general does not lead to selection consistency at the within-group level. In the proposed objective function (3.1.1), however, we adopt another

bridge penalty to induce within-group sparsity in replace of the L_1 penalty, and hence we refer to (3.1.1) as a composite group bridge penalized regression criterion. When $\mu = 1$, it reduces to the group bridge method; when $\gamma = 1$, it reduces to the form of a bridge regression. We shall mainly consider $\mu \in (0, 1)$ and $\gamma \in (0, 1)$. As we will show later, this leads to variable selection consistency at both the group and individual levels simultaneously. For simplicity and to streamline the idea, we fix $\gamma = \mu = 0.5$ in all our numerical studies. We note that the choice of μ or γ can also be made data adaptive, which may further boost the performance of the proposed approach.

3.2 Optimization

The minimization of the objective function (3.1.1) is challenging, as the composite group bridge penalty is nonconvex for $\mu \in (0, 1)$ and/or $\gamma \in (0, 1)$. Motivated by Huang *et al.* (2009), we show that an equivalent minimization problem can be formulated through an augmented variable approach, and an efficient iteratively reweighted Lasso regression algorithm is then developed for solving (3.1.1).

Define

$$S_{ln}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-\frac{1}{\gamma}} c_j^{\frac{1}{\gamma}} \left(\sum_{k \in A_j} \delta_k^{1-\frac{1}{\mu}} |\beta_k| + \psi \sum_{k \in A_j} \delta_k \right) + \tau_n \sum_{j=1}^{J} \theta_j, \tag{3.2.1}$$

where (τ_n, ψ) are some penalty parameters. The following proposition shows the equivalence between the minimizers of (3.2.1) and (3.1.1).

Proposition 3.2.1. *The composite group bridge estimator $\hat{\boldsymbol{\beta}}_n$ minimizes (3.1.1) if and only if*

$$(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}}) = \arg \min_{(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})} S_{ln}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}) \text{ subject to } \boldsymbol{\theta} \geq 0, \boldsymbol{\delta} \geq 0.$$

for some $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\delta}}$, where $\tau_n = \lambda_n^{1/(1-\gamma)} \gamma^{\gamma/(1-\gamma)} (1 - \gamma)$ and $\psi = \mu^{\mu/(1-\mu)} (1 - \mu)$.

It can be seen that by minimizing S_{ln} with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$ in (3.2.1), we can induce sparse solutions at both the group and individual levels, i.e., small θ_j forces $\boldsymbol{\beta}_{A_j} = 0$, leading to group selection, and small δ_k forces $\beta_k = 0$, leading to individual variable selection. For fixed $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, the problem becomes an adaptive Lasso problem in $\boldsymbol{\beta}$, which could be solved efficiently by many methods. Also, solving $\boldsymbol{\theta}$ with $(\boldsymbol{\beta}, \boldsymbol{\delta})$ held fixed and solving $\boldsymbol{\delta}$ with $(\boldsymbol{\beta}, \boldsymbol{\theta})$ held fixed both lead to explicit solutions. We therefore propose the following iterative algorithm for solving (3.2.1) and hence (3.1.1).

Composite Group Bridge Regression Algorithm

Initialization : start with an initial estimator $\boldsymbol{\beta}^{(0)}$, which can be obtained by least squares, Lasso or group Lasso methods.

For $s = 1, 2...,$

1. Calculate

$$\theta_j^{(s)} = c_j(\lambda_n\gamma)^{\frac{\gamma}{\gamma-1}}(\sum_{k\in A_j}|\beta_k^{(s-1)}|^\mu)^\gamma, \qquad j = 1, 2\ldots, J,$$
$$\delta_k^{(s)} = \mu^{\frac{\mu}{\mu-1}}|\beta_k^{(s-1)}|^\mu, \qquad k = 1, \ldots d.$$

2. Solve the adaptive Lasso problem,

$$\begin{aligned}\boldsymbol{\beta}^{(s)} &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \sum_{k=1}^{d}\mathbf{x}_k\beta_k\|_2^2 + \sum_{j=1}^{J}\{\theta_j^{(s)}\}^{1-\frac{1}{\gamma}}c_j^{\frac{1}{\gamma}}\sum_{k\in A_j}\{\delta_k^{(s)}\}^{1-\frac{1}{\mu}}|\beta_k| \\ &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \sum_{k=1}^{d}\mathbf{x}_k\beta_k\|_2^2 + \lambda_n\sum_{k=1}^{d}w_{1k}^{(s)}|\beta_k|,\end{aligned}$$

where

$$w_{1k}^{(s)} = \gamma\mu\sum_{j:k\in A_j}c_j\|\boldsymbol{\beta}_{A_j}^{(s-1)}\|_\mu^{\mu(\gamma-1)}|\beta_k^{(s-1)}|^{\mu-1}. \tag{3.2.2}$$

3. Repeat steps 1–2 until convergence, e.g., $\|\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s-1)}\|_2/\|\boldsymbol{\beta}^{(s-1)}\|_2 < \epsilon$, where ϵ is some tolerance level, e.g. $\epsilon = 10^{-4}$.

The proposed algorithm has a blockwise coordinate descent structure, in which we alternatively update $\boldsymbol{\theta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, one block at a time with other two blocks held fixed. It is evident that the proposed composite group bridge method is closely related to adaptive Lasso, as the above algorithm boils down to an iteratively reweighted adaptive Lasso procedure. Another way to reveal this connection is from adopting a local linear approximation (Zou and Li, 2008) of the composite group bridge penalty in (3.1.1). Suppose $\boldsymbol{\beta}^{(0)}$ be an initial estimator of $\boldsymbol{\beta}_0$, and denote $\boldsymbol{\beta}^{(0)}_{-l}$ as a subvector of $\boldsymbol{\beta}^{(0)}$ without its lth entry $\boldsymbol{\beta}^{(0)}_l$, for any $l = 1, \ldots, d$. For fixed $\boldsymbol{\beta}^{(0)}_{-l}$, consider the penalty terms involving the lth predictor,

$$\begin{aligned} f(\beta_l; \boldsymbol{\beta}^{(0)}_{-l}) =& \sum_{j: l \in A_j} c_j \Big(\sum_{k \in A_j, k \neq l} |\beta^{(0)}_k|^\mu + |\beta_l|^\mu \Big)^\gamma \\ \approx & f(\beta^{(0)}_l; \boldsymbol{\beta}^{(0)}_{-l}) + f'(\beta^{(0)}_l; \boldsymbol{\beta}^{(0)}_{-l}) \{\beta_l - \beta^{(0)}_l\} \\ = & f(\beta^{(0)}_l; \boldsymbol{\beta}^{(0)}_{-l}) + \left[\gamma\mu \sum_{j: l \in A_j} c_j \|\boldsymbol{\beta}^{(0)}_{A_j}\|_\mu^{\mu(\gamma-1)} |\beta^{(0)}_l|^{\mu-1} \right] \{\beta_l - \beta^{(0)}_l\} \end{aligned}$$

It can be seen that for fixed $\boldsymbol{\beta}^{(0)}$, up to a constant, the first-order approximation yields exactly an adaptive Lasso penalty for β_l, and the weight takes exactly the same form as (3.2.2). This provides an alternative justification of the validity of the proposed algorithm.

As the objective function decreases monotonically along the updates, the convergence of the algorithm is guaranteed. However, due to the nonconvexity of the proposed criterion, the algorithm in general converges to a local minimizer. Based on our limited experience, the proposed method is stable and performs well in practice.

For any fixed $\lambda \geq 0$, the minimizer of (3.1.1) can be computed by the preceding algorithm. To choose an optimal λ and hence an optimal solution, a general method is to use K-fold cross validation (CV) (Stone, 1974). However, using CV can be computationally expensive for large data. We mainly use an BIC criterion (Schwarz, 1978) for tuning because of its computational efficiency and promising performance on sparse estimation. Denote $\hat{\boldsymbol{\beta}}(\lambda)$ as the estimator of $\boldsymbol{\beta}_0$ by solving (3.1.1). We define

$$\text{BIC}(\lambda) = \log \Big[RSS\{\hat{\boldsymbol{\beta}}(\lambda)\}/n \Big] + \log\{\max(p, n)\} df(\lambda)/n, \tag{3.2.3}$$

where $RSS\{\hat{\boldsymbol{\beta}}(\lambda)\}$ is defined in (2.1.2) and $df(\lambda)$ is the effective degrees of freedom of the fitted model. Because of the iterative adaptive Lasso interpretation of the final estimator, we have used the number of nonzero coefficients to estimate the model degrees of freedom, $\hat{df}(\lambda) = \|\hat{\boldsymbol{\beta}}(\lambda)\|_0$, following Zou *et al.* (2007) and Breheny and Huang (2009a). In our numerical studies, we compute the solutions over a grid of 100 λ values equally spaced on the log scale and then select the best λ value by BIC.

3.3 Adaptive Estimators

From the preceding discussion, it is evident that the proposed composite group bridge method is closely related to adaptive Lasso. Similar to Zou and Li (2008), a one-step adaptive Lasso estimator can be readily constructed with the weights take the form as in (3.2.2). As the adaptive Lasso is a convex problem, the one-step estimator is a global minimizer, easy to compute, and its theoretical properties are readily available (Zou, 2006; Huang *et al.*, 2008b). However, there are some drawbacks. The performance of adaptive estimation relies heavily on the quality of the initial estimator used in constructing the adaptive weights. Moreover, adaptive Lasso essentially is an individual variable selection tool; although the weights (3.2.2) have a group-level component, empirical study suggests that the group level selection by the adaptive Lasso method is often unsatisfactory.

As a compromise, we propose an adaptive group bridge criterion,

$$L_n(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_{1,w}^{\gamma}, \tag{3.3.1}$$

where $\|\cdot\|_{1,w}$ denotes the $\mathbf{w}$-weighted L_1 norm for the enclosed vector, e.g., $\|\boldsymbol{\beta}_{A_j}\|_{1,w} = \sum_{k\in A_j} w_k|\beta_k|$, w_ks are some individual-level weights, c_js are some group-level weights, and all the other terms are similarly defined as in (3.1.1).

Suppose some reliable initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ is available, e.g. the least squares estimator when the sample size is large relative to the model dimension. We set

$$w_k = \frac{|\hat{\beta}_k^{(0)}|^{-\mu}}{\alpha_k},$$

where $\alpha_k = \sum_j I(k \in A_j)$ counts the number of groups the kth predictor belongs to, and $\mu \geq 0$ is a power parameter usually set to be 2. For the group-level weights, we can simply set

$$c_j = |A_j|^{1-\gamma}, \tag{3.3.2}$$

accounting for the group size. When $\mu > 1$, we also consider

$$c_j = \left(\sum_{k \in A_j} \frac{|\hat{\beta}_k^{(0)}|^{1-\mu}}{\alpha_k} \right)^{1-\gamma}, \tag{3.3.3}$$

adjusting for the magnitude of each group of weighted coefficients.

It is straightforward to show that the problem can also be solved by the proposed iterative adaptive Lasso algorithm. The only change is that in step 3, the weights become

$$w_{1k}^{(s)} = \gamma \sum_{j:k \in A_j} c_j \|\boldsymbol{\beta}_{A_j}^{(s-1)}\|_{1,w}^{\gamma-1} w_k. \tag{3.3.4}$$

Comparing to composite group bridge, the problem is simpler and remains to possess good properties when reliable adaptive weights are available.

Chapter 4

Asymptotic Properties

In this section, we explore the asymptotic properties of the proposed composite group bridge estimator. We show that, for $0 < \gamma, \mu < 1$, the composite group bridge estimator identifies the correct groups and the correct nonzero coefficients within each selected group with probability converging to one, under reasonable conditions. The estimation error bound and asymptotic distribution of the proposed estimator are also established.

Recall that without loss of generality, we have assumed that

$$\boldsymbol{\beta}_{0A_j} \neq \mathbf{0}, 1 \leq j \leq J_1,$$
$$\boldsymbol{\beta}_{0A_j} = \mathbf{0}, J_1 + 1 \leq j \leq J.$$

For each A_j, $j = 1, ..., J_1$, $A_j^1 = \{k; \beta_{0k} \neq 0, k \in A_j\}$ and $A_j^2 = \{k; \beta_{0k} = 0, k \in A_j\}$.

Let $B_2 = \cup_{j=J_1+1}^{J} A_j$ be the union of the groups with zero coefficients. Let $B_1 = B_2^c$, $B_1^1 = \{k; \beta_{0k} \neq 0, k \in B_1\}$ and $B_1^2 = \{k; \beta_{0k} = 0, k \in B_1\}$. Note that each A_j may include important predictors, unimportant predictors in B_1 and unimportant predictors in B_2. Denote $\boldsymbol{\beta}_{0B_j} = (\beta_{0k}, k \in B_j)'$ for $j = 1, 2$, and define other subvectors of $\boldsymbol{\beta}_0$ similarly. Assume the variables are arranged so that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{0B_1^1}, \boldsymbol{\beta}'_{0B_1^2}, \boldsymbol{\beta}'_{0B_2})'$. Since $\boldsymbol{\beta}_{0B_1^2} = 0$ and $\boldsymbol{\beta}_{0B_2} = 0$, the response variable is fully explained by the important variables belonging to B_1^1 within the first J_1 groups. In this notation, $\hat{\boldsymbol{\beta}}_{nB_1^1}$, $\hat{\boldsymbol{\beta}}_{nB_1^2}$ and $\hat{\boldsymbol{\beta}}_{nB_2}$ are respectively the estimates of $\boldsymbol{\beta}_{0B_1^1}$, $\boldsymbol{\beta}_{0B_1^2}$ and $\boldsymbol{\beta}_{0B_2}$ from the composite group bridge estimator $\hat{\boldsymbol{\beta}}_n$.

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d)$, $\mathbf{X}_1 = (\mathbf{x}_k, k \in B_1)$, $\mathbf{X}_{11} = (\mathbf{x}_k, k \in B_1^1)$ and $\mathbf{X}_{12} = (\mathbf{x}_k, k \in B_1^2)$.

Define

$$\mathbf{\Sigma}_n = \frac{1}{n}\mathbf{X}'\mathbf{X}, \qquad \mathbf{\Sigma}_{1n} = \frac{1}{n}\mathbf{X}_1'\mathbf{X}_1, \qquad \mathbf{\Sigma}_{11n} = \frac{1}{n}\mathbf{X}_{11}'\mathbf{X}_{11} \tag{4.0.1}$$

Let ρ_n and ρ_n^* be the smallest and largest eigenvalues of $\mathbf{\Sigma}_n$, and let τ_{1n} and τ_{1n}^* be the smallest and largest eigenvalues of $\mathbf{\Sigma}_{11n}$.

We consider the following conditions.

A1. The errors $\varepsilon_1, \varepsilon_2, ...\varepsilon_n$ are uncorrelated with mean zero and finite variance σ^2.

A2. The maximum multiplicity $c_n^* = \max_k \sum_{j=1}^J I(k \in A_j)$ is bounded, and

$$\frac{\lambda_n^2 \eta_n^2}{n\rho_n \sigma^2 d} = M_n = O(1),$$

where $\eta_n = \{\sum_{j=1}^{J_1} c_j^2 \|\boldsymbol{\beta}_{0A_j^1}\|_{2\mu-2}^{2\mu-2} \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{2\mu(\gamma-1)}\}^{1/2}$.

A3. The constants c_js are scaled to satisfy $\min_{j\le J} c_j \ge 1$, and

$$\frac{\lambda_n \rho_n^{1-\mu\gamma/2}}{d^{1-\mu\gamma/2}\rho_n^* n^{\mu\gamma/2}} \to \infty.$$

A3*. The constants c_js are scaled to satisfy $\min_{j\le J} c_j \ge 1$, and

$$\frac{\lambda_n \rho_n^{1-\mu/2}}{d^{1-\mu/2}\rho_n^* n^{\mu/2}} \to \infty.$$

A4. There exist constants $\tau_1^* < \infty$ such that $\tau_{1n}^* \le \tau_1^*$ for all n.

Assumption A1 is standard about the error distribution. Assumptions A2 and A3 are about the degree of overlapping, the growth rate of the tuning parameter and the growth rate of the model size; they imply full rank design with $\text{rank}(\mathbf{X}) = d \le n$, $\rho_n > 0$, and $\tau_{1n} > 0$. The first three assumptions are used to establish the group level selection consistency. To establish individual level selection consistency, however, A3 shall be replaced by a stronger version A3*, and A4 is also needed to ensure the largest eigenvalue of $\mathbf{X}_{11}$ is bounded.

In general, the selection consistency of the individual level is stronger than that of the group level. This fact is reflected in the above required assumptions. Note that A3* implies

A3, as individual level selection consistency implies group level consistency. The choice of the group level penalty does not have a direct impact on the within-group variable selection. On the other hand, A3 involves both μ and γ, because the choice of the within-group penalty determines the behavior of individual variable selection and hence also influences the group selection performance. Similar to Huang *et al.* (2009), for $[B_1^1, \boldsymbol{\beta}_{0B_1^1}, J_1]$ fixed but unknown, assumptions A2 and A3 hold when

$$\begin{aligned} &\text{(a) } (1/\rho_n) + \rho_n^* + \sum_{j=1}^{J_1} c_j^2 = O(1), \\ &\text{(b) } \frac{\lambda_n}{n^{1/2}} \to \lambda_0 < \infty, \\ &\text{(c) } \frac{\lambda_n d^{\mu\gamma/2}}{d n^{\mu\gamma/2}} \to \infty, \end{aligned} \tag{4.0.2}$$

provided that $c_j \geq 1$ and $c_n^* = O(1)$. For assumptions A2 and A3*, (c) is strengthened to

$$\text{(c*) } \frac{\lambda_n d^{\mu/2}}{d n^{\mu/2}} \to \infty. \tag{4.0.3}$$

Here the number of covariates $d = d_n$ is allowed to grow at a certain rate as $n > d_n \to \infty$. To ensure group selection consistency, we allow $d_n = o(1)n^{(1-\mu\gamma)/(2-\mu\gamma)}$, which is faster than what is allowed by group bridge. To ensure both group and individual level selection consistency, we allow a slower rate $d_n = o(1)n^{(1-\mu)/(2-\mu)}$.

Theorem 4.0.1 (Estimation Error Bound)**.** *Suppose that* $0 < \mu \leq 1$, $0 < \gamma < 1$ *and assumptions A1–A2 hold. Then*

$$E(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2) \leq \frac{\sigma^2 d}{n\rho_n}(8 + 64 c_n^* M_n).$$

Theorem 4.0.2 (Group Selection Consistency)**.** *Suppose that* $0 < \mu \leq 1$, $0 < \gamma < 1$ *and assumptions A1–A3 hold. Then,*

$$P(\hat{\boldsymbol{\beta}}_{nA_j} = 0, j > J_1) \to 1, \tag{4.0.4}$$

as $n \to \infty$.

Theorem 4.0.3 (Individual Selection Consistency)**.** *Suppose that* $0 < \mu < 1$, $0 < \gamma \leq 1$ *and assumptions A1, A2, A3* and A4 hold. Then (4.0.4) holds, and*

$$P(\hat{\boldsymbol{\beta}}_{nA_j^2} = 0, j \leq J_1) \to 1,$$

as $n \to \infty$.

Theorem 4.0.4 (Asymptotic Distribution)**.** *Suppose* $\{B_1^1, \boldsymbol{\beta}_{0B_1^1}, J_1\}$ *are fixed unknowns and (4.0.2) holds. Suppose further that* $\boldsymbol{\Sigma}_{1n} \to \boldsymbol{\Sigma}_1$ *and* $n^{-1/2}\mathbf{X}_1'\boldsymbol{\varepsilon} \to \mathbf{W}_1 \sim N(0, \sigma^2\boldsymbol{\Sigma}_1)$, *and consequently* $\boldsymbol{\Sigma}_{11n} \to \boldsymbol{\Sigma}_{11}$ *and* $n^{-1/2}\mathbf{X}_{11}'\boldsymbol{\varepsilon} \to \mathbf{W}_{11} \sim N(0, \sigma^2\boldsymbol{\Sigma}_{11})$. *Then,*

$$\sqrt{n}\hat{\boldsymbol{\beta}}_{nB_2} \to_d \boldsymbol{0}, \sqrt{n}\hat{\boldsymbol{\beta}}_{nB_1^2} \to_d \boldsymbol{0},$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{nB_1^1} - \boldsymbol{\beta}_{0B_1^1}) \to_d \arg\min V_{11}(\mathbf{u}), \mathbf{u} \in \mathbb{R}^{|B_1^1|},$$

where

$$V_{11}(\mathbf{u}) = -2\mathbf{u}'\mathbf{W}_{11} + \mathbf{u}'\boldsymbol{\Sigma}_{11}\mathbf{u} + \mu\gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\boldsymbol{\beta}_{0A_j}\|_\mu^{\mu(\gamma-1)} \sum_{k \in A_j^1} u_k |\beta_{0k}|^{\mu-1} sgn(\beta_{0k}).$$

Chapter 5

Simulation

5.1 Simulation Setup

We compare the proposed composite group bridge (CoGB) and adaptive group bridge (AdGB) techniques with group bridge (GB) (Huang *et al.*, 2009), composite MCP (CoMCP) (Breheny and Huang, 2009b), group Lasso (grLasso) (Yuan and Lin, 2006), group MCP (grMCP) (Breheny, 2009), and adaptive Lasso (AdLasso) (Zou, 2006) methods for variable selection. The grLasso, grMCP, GB and CoMCP estimators are computed using the *grpreg* package (Breheny and Huang, 2009b) in R (R Development Core Team, 2008). We have also implemented all the other methods considered in this report in R. We have experimented with several information criteria including BIC, AIC and GCV for tuning parameter selection, and in general BIC gives better results on variable selection. Hence we shall compare all the methods with BIC tuning.

We consider five simulation examples covering various practical scenarios, e.g., bi-level sparsity, group sparsity, overlapping of predictors, varying group sizes, correlation within and among groups, etc. The setup is in a similar fashion as the simulation study in Huang *et al.* (2009). We consider sample sizes of $n = 100$, 200 and 400. The experiment is replicated 400 times under each setting.

Example 1: There are $J = 6$ groups of variables, with $|A_1| = |A_2| = |A_3| = 10$ and

$|A_4| = |A_5| = |A_6| = 4$. To generate $d = 42$ covariates, we first form $n \times 1$ vectors $\mathbf{r}_1, \ldots, \mathbf{r}_d$ and $\mathbf{z}_1, \ldots, \mathbf{z}_J$; all the entries in these vectors are independently generated from $N(0, 1)$. The covariates $\mathbf{x}_1, .., \mathbf{x}_d$ are generated as

$$\mathbf{x}_k = (\mathbf{z}_{g_k} + \mathbf{r}_k)/\sqrt{2}, \qquad 1 \le k \le d,$$

where $(g_1, \ldots, g_d)$=$(\underbrace{1, ..., 1}_{10}, \underbrace{2, ..., 2}_{10}, \underbrace{3, ..., 3}_{10}, \underbrace{4, ..., 4}_{4}, \underbrace{5, ..., 5}_{4}, \underbrace{6, ..., 6}_{4})$, indicating the group membership structure. Therefore, the covariates within each group are correlated, while the covariates from different groups are uncorrelated. The response $\mathbf{y}$ is then generated using model (2.1.1), where

$$\boldsymbol{\beta}_{0A_1} = (1, -2, 1.25, 1, -1, 1, 3, -1.5, 2, -2)', \qquad \boldsymbol{\beta}_{0A_2} = (-1.5, 3, 1, -2, 1.5, 0, 0, 0, 0, 0)',$$

$$\boldsymbol{\beta}_{0A_3} = (0, \ldots, 0)', \quad \boldsymbol{\beta}_{0A_4} = (2, -2, 1, 1.5)', \quad \boldsymbol{\beta}_{0A_5} = (-1.5, 1.5, 0, 0)', \quad \boldsymbol{\beta}_{0A_6} = (0, \ldots, 0)',$$

and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 4\mathbf{I})$.

Example 2: The model is the same as in Example 1, except that

$$\boldsymbol{\beta}_{0A_2} = (-1.5, 3, 0, \ldots, 0)'.$$

So the second group is very sparse.

Example 3: We consider overlapping of predictors within groups. There are six groups, with $|A_1| = 10$, $|A_2| = 11$, $|A_3| = 12$, $|A_4| = 4$, $|A_5| = 4$, and $|A_6| = 4$. The 10th covariate belongs to both the first and the second groups, and the 19th and 20th covariates belong to both the second and the third groups, so the total number of covariates still equals to $d = 42$. The covariates $\mathbf{x}_1$,...,$\mathbf{x}_{42}$ are generated in the same way as in Example 1, except that for $k = 10$, $\mathbf{x}_k = (\mathbf{z}_1 + \mathbf{z}_2 + \mathbf{r}_k)/\sqrt{3}$, and for $k = 19, 20$, $\mathbf{x}_k = (\mathbf{z}_2 + \mathbf{z}_3 + \mathbf{r}_k)/\sqrt{3}$. Therefore, each overlapping covariate is correlated with all the other covarites from the groups it belongs to. The response is computed using model (2.1.1), where $\boldsymbol{\beta}_0$ is exactly the same as

in Example 1 and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 4\mathbf{I})$.

Example 4: We consider a very sparse model. In this example, $J = 5$, $|A_1| = \cdots = |A_5| = 8$ and $d = 40$, so there is no overlap of group membership. The covariates $\mathbf{x}_1$,...,$\mathbf{x}_{40}$ are generated the same way as in Example 1. The response $\mathbf{y}$ is generated using model (2.1.1), where

$$\boldsymbol{\beta}_{0A_1} = (0, 0, 0, 2, 0, 2, 0, 0)', \qquad \boldsymbol{\beta}_{0A_2} = (3, 3, 0, 0, 0, 0, 0, 0, 0, 0)',$$

$\boldsymbol{\beta}_{0A_3} = \boldsymbol{\beta}_{0A_4} = \boldsymbol{\beta}_{0A_5} = \mathbf{0}'$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 4\mathbf{I})$.

Example 5: We consider a model in which all the coefficients in a nonzero group are nonzero, with $J = 5$, $|A_1| = \cdots = |A_5| = 8$ and $d = 40$. We first simulate $\mathbf{r}_1, ..., \mathbf{r}_{40}$ independently from $N(\mathbf{0}, \mathbf{I})$. Next, to generate $\mathbf{z}_j$ vectors ($j = 1, \ldots, J$), we simulate n independent samples from a J-dimensional Gaussian distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the (h, l)th entry of $\boldsymbol{\Sigma}$ equals to $\sigma_{hl} = 0.4^{|h-l|}$. Then the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_{40}$ are generated as

$$\mathbf{x}_{5(j-1)+k} = \{\mathbf{z}_j + \mathbf{r}_{4(j-1)+k}\}/\sqrt{2}. \qquad 1 \le j \le 5, \quad 1 \le k \le 8.$$

In this way, the AR(1) correlation structure of the $\mathbf{z}_j$s induces correlation across different groups of covariates. The response vector is computed using model (2.1.1), where

$$\boldsymbol{\beta}_{0A_1} = (1, 1, 1.5, 2, 2.5, 3, 3.5, 4)', \qquad \boldsymbol{\beta}_{0A_2} = (2, 2, 2, 2, 2, 2, 2, 2)',$$

$\boldsymbol{\beta}_{0A_3} = \boldsymbol{\beta}_{0A_3} = \boldsymbol{\beta}_{0A_3} = \mathbf{0}'$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 4\mathbf{I})$.

5.2 Evaluation Methods

For each method, the model accuracy is measured by the average of the model error from all 400 runs (Model Error), i.e., $(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})$ where $\boldsymbol{\Sigma}$ is the true covariance matrix of the predictors. To evaluate the sparse bi-level estimation performance, we report the

average number of nonzero coefficients (No. of Var.) and the average number of selected nonzero groups (No. of Groups). To further evaluate bi-level variable selection accuracy, we report the following measures,

1. the frequency of correct identification of the group sparsity structure (Correct Groups),
2. the frequency of correct identification of the bi-level sparsity structure (Correct Model),
3. the false negative rate (FNR) of missing important predictors,
4. the false discovery rate (FDR) of selecting unimportant predictors.

To be precise, for each fitted model, denote the number of false positives as FP, the number of false negatives as FN, the number of true positives as TP and number of true negatives as TN. Then FNR is calculated as FN/(TP+FN), and FDR is calculated as FP/(FP+TP). Tables 1–3 summarize the simulation results for sample sizes 400, 200 and 100, respectively.

5.3 Simulation Results

The first three examples are in favor of the bi-level selection methods. In terms of estimation accuracy, AdGB, CoGB and CoMCP have comparable performance, and they slightly outperform GB, which is also capable of conducting bi-level selection but does not enjoy individual level selection consistency. All these methods generally outperform grMCP, grLasso and AdLasso, which do not perform bi-level selection, and the advantage of the former set of methods is persistent across different sample sizes. Not surprisingly, AdGB, CoGB and CoMCP all perform very well in variable selection, and the FNR and FPR rates are general low. CoMCP slightly outperforms CoGB, mainly because the latter tends to have a slightly larger false negative rate than the former. We also note that grMCP also performs much better than grLasso in group selection and model estimation. These show the superior performance of a well designed nonconvex MCP penalty to the convex Lasso penalty and the nonconvex but simpler bridge penalty. Nevertheless, CoGB performs much better than GB

in terms of correct model identification, supporting the developed bi-level selection oracle property.

Example 4 is designed in favor of AdLasso, as the model is extremely sparse. As expected, AdLasso performs better than most of the methods. As all the nonzero groups are very sparse, i.e, very few important predictors are grouped with a large number of irrelevant predictors, we may view this structure as a case of severe group misidentification. Our results show that the bi-level selection methods still perform satisfactorily, while grMCP and grLasso perform poorly in model estimation because of high false positives. This demonstrates that using bi-level selection methods may greatly alleviate the influence of group misspecification, and in practice it is especially beneficial when the prior group information may be unreliable. In Example 5, the model is sparse at the group level but nonsparse at the within group level. Therefore, this is exactly the structure the group selection methods aim to recover. As expected, both grMCP and grLasso perform extremely well in model selection. But again, the bi-level selection methods are not far behind.

To further examine the variable selection performance of various methods, for each covariate in a simulated model, we plot the relative frequency that it is not selected in the 400 simulation runs. The plots for the first three examples with sample sizes 400, 200 and 100 are shown in Figures 2, 3 and 4. CoGB estimates are represented as solid blue dots, AdGB estimates are represented as filled red squares, CoMCP estimates are represented as purple stars, and GB estimates are represented as green triangle point-ups. It is clear that GB method often yields false positives, and the other three bi-level selection methods perform much better because of their enhanced individual level selection capability.

Table 5.1: *Comparison of various estimators when the sample size is 400. The standard errors are reported in parentheses. CoGB: composite group bridge; AdGB(1): adaptive group bridge estimator with group weights in (3.3.3); AdGB(2): adaptive group bridge estimator with group weights in (3.3.2); GB: group bridge; CoMCP: composite MCP; grMCP: group MCP; grLasso: group Lasso; AdLasso; adaptive Lasso.*

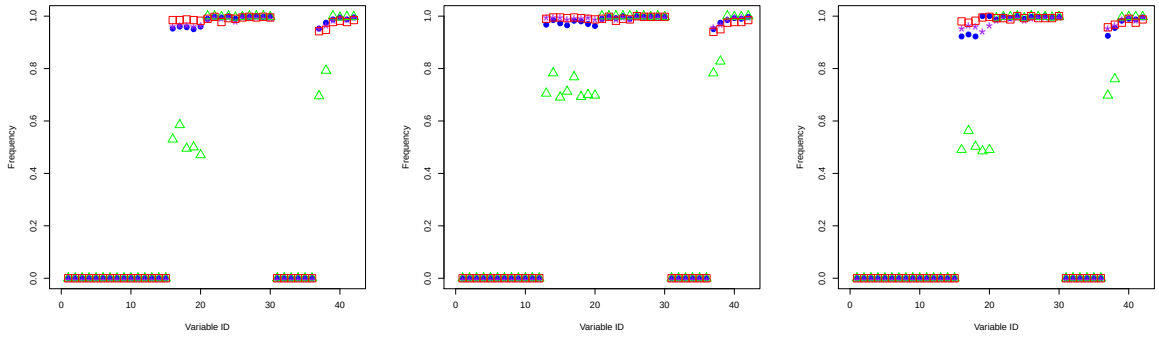
Measure	True Model	CoGB	AdGB(1)	AdGB(2)	CoMCP	GB	grMCP	grLasso	AdLasso
					Example 1				
Model Error	0.00	0.25 (0.08)	0.24 (0.08)	0.24 (0.08)	0.25 (0.08)	0.34 (0.11)	0.30 (0.08)	0.57 (0.16)	0.30 (0.09)
No of Vars	21.00	21.36 (0.69)	21.34 (0.95)	21.26 (0.57)	21.31 (0.70)	23.96 (1.62)	28.00 (0.00)	33.38 (5.90)	21.83 (1.02)
No of Groups	4.00	4.06 (0.26)	4.10 (0.34)	4.12 (0.35)	4.07 (0.28)	4.02 (0.16)	4.00 (0.00)	4.84 (0.83)	4.50 (0.63)
Correct Group(%)	100.00	94.25	90.75	88.75	94.50	98.25	100.00	43.75	58.25
Correct Model(%)	100.00	73.50	78.00	80.00	78.00	5.50	0.00	0.00	47.75
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	1.59	1.42	1.13	1.36	11.94	25.00	35.23	3.60
					Example 2				
Model Error	0.00	0.22 (0.07)	0.21 (0.07)	0.21 (0.07)	0.20 (0.07)	0.33 (0.11)	0.30 (0.08)	0.57 (0.17)	0.26 (0.09)
No of Vars	18.00	18.36 (0.69)	18.32 (0.96)	18.25 (0.58)	18.14 (0.43)	20.65 (1.72)	28.00 (0.00)	33.38 (5.94)	18.72 (0.93)
No of Groups	4.00	4.07 (0.26)	4.11 (0.34)	4.12 (0.34)	4.05 (0.22)	4.01 (0.11)	4.00 (0.00)	4.84 (0.83)	4.39 (0.56)
Correct Group(%)	100.00	93.50	90.50	88.75	95.75	99.50	100.00	44.25	64.75
Correct Model(%)	100.00	73.25	78.75	81.00	88.75	8.00	0.00	0.00	51.25
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	1.84	1.55	1.29	0.73	12.26	35.71	44.46	3.61
					Example 3				
Model Error	0.00	0.26 (0.08)	0.24 (0.08)	0.24 (0.08)	0.25 (0.08)	0.33 (0.10)	0.29 (0.08)	0.60 (0.18)	0.30 (0.09)
No of Vars	21.00	21.43 (0.73)	21.28 (0.93)	21.25 (0.56)	21.34 (0.73)	24.06 (1.67)	28.00 (0.00)	33.67 (0.36)	21.89 (1.09)
No of Groups	4.00	4.08 (0.31)	4.09 (0.33)	4.13 (0.36)	4.14 (0.37)	4.77 (0.44)	5.00 (0.00)	5.52 (0.50)	4.54 (0.64)
Correct Group(%)	100.00	92.50	92.75	87.75	87.50	24.00	0.00	0.00	53.50
Correct Model(%)	100.00	68.25	83.25	80.50	77.25	4.50	0.00	0.00	46.25
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	1.91	1.15	1.11	1.47	12.30	25.00	35.69	3.85
					Example 4				
Model Error	0.00	0.06 (0.04)	0.06 (0.06)	0.06 (0.05)	0.04 (0.03)	0.12 (0.07)	0.17 (0.06)	0.43 (0.16)	0.05 (0.04)
No of Vars	4.00	4.29 (0.69)	4.28 (0.87)	4.30 (0.70)	4.06 (0.05)	5.98 (1.41)	16.00 (0.00)	16.08 (0.79)	4.14 (0.48)
No of Groups	2.00	2.07 (0.28)	2.12 (0.36)	2.16 (0.44)	2.03 (0.17)	2.00 (0.00)	2.00 (0.00)	2.01 (0.10)	2.10 (0.35)
Correct Group(%)	100.00	94.25	90.00	86.75	97.75	100.00	100.00	99.00	91.75
Correct Model(%)	100.00	80.75	83.25	80.25	95.25	13.00	0.00	0.00	89.50
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	5.01	4.44	5.18	1.11	29.55	75.00	75.08	2.55
					Example 5				
Model Error	0.00	0.18 (0.07)	0.18 (0.07)	0.18 (0.07)	0.17 (0.06)	0.17 (0.06)	0.17 (0.06)	0.37 (0.14)	0.19 (0.07)
No of Vars	16.00	16.26 (0.55)	16.18 (0.47)	16.23 (0.52)	16.09 (0.42)	16.09 (0.31)	16.00 (0.00)	16.08 (0.80)	16.19 (0.48)
No of Groups	2.00	2.24 (0.49)	2.16 (0.39)	2.20 (0.45)	2.08 (0.33)	2.08 (0.27)	2.00 (0.00)	2.01 (0.10)	2.19 (0.46)
Correct Group(%)	100.00	79.25	85.50	81.75	93.25	92.25	100.00	99.00	84.00
Correct Model(%)	100.00	79.25	85.50	81.75	93.25	92.25	100.00	99.00	84.00
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	1.48	1.02	1.29	0.50	0.49	0.00	0.33	1.11

Table 5.2: *Comparison of various estimators when the sample size is 200. All the settings are the same as in Table 5.1.*

Measure	True Model	CoGB	AdGB(1)	AdGB(2)	CoMCP	GB	grMCP	grLasso	AdLasso
					Example 1				
Model Error	0.00	0.58 (0.20)	0.56 (0.22)	0.57 (0.22)	0.59 (0.21)	0.69 (0.21)	0.65 (0.19)	1.22 (0.36)	0.71 (0.25)
No of Vars	21.00	21.72 (1.03)	21.65 (1.15)	21.60 (1.00)	21.68 (0.99)	24.42 (1.60)	28.00 (0.00)	35.43 (6.35)	22.89 (1.70)
No of Groups	4.00	4.13 (0.37)	4.16 (0.39)	4.22 (0.46)	4.15 (0.40)	4.03 (0.18)	4.00 (0.00)	5.10 (0.88)	4.89 (0.69)
Correct Group(%)	100.00	88.00	85.00	80.50	87.00	96.75	100.00	34.25	29.75
Correct Model(%)	100.00	53.75	63.00	61.00	58.50	3.25	0.00	0.00	21.25
FNR	0.00	0.02	0.06	0.11	0.02	0.00	0.00	0.00	0.06
FDR	0.00	3.15	2.82	2.67	2.96	13.62	25.00	38.73	7.84
					Example 2				
Model Error	0.00	0.48 (0.18)	0.46 (0.18)	0.48 (0.18)	0.46 (0.16)	0.66 (0.22)	0.65 (0.19)	1.22 (0.37)	0.62 (0.22)
No of Vars	18.00	18.65 (1.06)	18.57 (1.16)	18.53 (0.93)	18.36 (0.72)	21.24 (1.87)	28.00 (0.00)	35.25 (6.37)	19.76 (1.57)
No of Groups	4.00	4.14 (0.39)	4.16 (0.39)	4.20 (0.45)	4.11 (0.34)	4.01 (0.09)	4.00 (0.00)	5.07 (0.87)	4.77 (0.67)
Correct Group(%)	100.00	87.75	85.50	81.75	90.50	99.25	100.00	35.00	36.25
Correct Model(%)	100.00	62.50	69.00	64.25	75.75	5.25	0.00	0.00	22.75
FNR	0.00	0.01	0.00	0.08	0.00	0.00	0.00	0.00	0.08
FDR	0.00	3.20	2.76	2.74	1.83	14.62	35.71	47.22	8.45
					Example 3				
Model Error	0.00	0.59 (0.19)	0.55 (0.22)	0.57 (0.21)	0.59 (0.21)	0.69 (0.22)	0.64 (0.18)	1.21 (0.37)	0.70 (0.23)
No of Vars	21.00	21.86 (1.08)	21.64 (1.61)	21.53 (0.98)	21.66 (1.01)	24.43 (1.78)	28.00 (0.00)	36.53 (6.16)	22.93 (1.70)
No of Groups	4.00	4.20 (0.47)	4.16 (0.43)	4.23 (0.47)	4.23 (0.49)	4.79 (0.45)	5.00 (0.00)	5.66 (0.47)	4.95 (0.68)
Correct Group(%)	100.00	83.50	86.50	79.50	76.75	22.50	0.00	0.00	25.75
Correct Model(%)	100.00	46.50	68.25	65.00	55.25	3.00	0.00	0.00	18.50
FNR	0.00	0.01	0.06	0.10	0.06	0.00	0.00	0.00	0.06
FDR	0.00	3.71	2.63	2.38	2.93	13.58	25.00	40.68	8.02
					Example 4				
Model Error	0.00	0.12 (0.09)	0.12 (0.10)	0.19 (0.09)	0.09 (0.07)	0.23 (0.14)	0.36 (0.14)	0.87 (0.36)	0.11 (0.08)
No of Vars	4.00	4.46 (0.81)	4.36 (0.77)	4.41 (0.81)	4.12 (0.42)	6.26 (1.51)	16.00 (0.00)	16.16 (1.12)	4.22 (0.59)
No of Groups	2.00	2.11 (0.34)	2.14 (0.38)	2.18 (0.43)	2.05 (0.23)	2.00 (0.00)	2.00 (0.00)	2.02 (0.14)	2.14 (0.39)
Correct Group(%)	100.00	90.00	87.00	84.00	95.25	100.00	100.00	98.00	87.75
Correct Model(%)	100.00	69.25	76.25	73.00	90.75	11.00	0.00	0.00	84.00
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	8.02	6.10	7.06	2.23	32.33	75.00	75.16	3.96
					Example 5				
Model Error	0.00	2.32 (0.54)	0.18 (0.07)	0.42 (0.17)	0.36 (0.14)	0.36 (0.14)	0.35 (0.13)	0.77 (0.31)	0.49 (0.20)
No of Vars	16.00	16.39 (0.72)	16.18 (0.47)	16.31 (0.69)	16.14 (0.57)	16.19 (0.53)	16.00 (0.00)	16.04 (0.56)	16.45 (0.88)
No of Groups	2.00	2.32 (0.54)	2.22 (0.47)	2.26 (0.52)	2.12 (0.39)	2.15 (0.36)	2.00 (0.00)	2.01 (0.07)	2.40 (0.65)
Correct Group(%)	100.00	72.00	80.25	77.25	90.50	85.25	100.00	99.50	68.75
Correct Model(%)	100.00	72.00	80.00	76.50	90.50	85.25	100.00	99.50	67.25
FNR	0.00	0.00	0.02	0.05	0.00	0.00	0.00	0.00	0.09
FDR	0.00	2.19	1.51	1.76	0.74	1.09	0.00	0.17	2.57

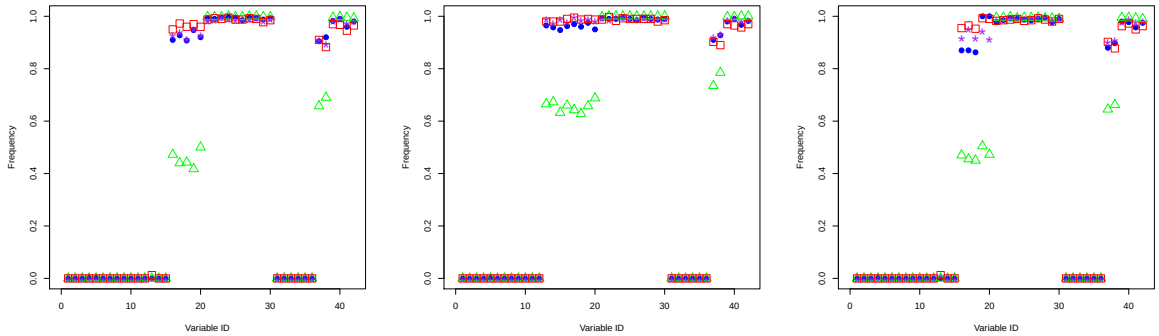
Table 5.3: *Comparison of various estimators when the sample size is 100. All the settings are the same as in Table 5.1.*

Measure	True Model	CoGB	AdGB(1)	AdGB(2)	CoMCP	GB	grMCP	grLasso	AdLasso
					Example 1				
Model Error	0.00	1.70 (0.70)	1.69 (0.78)	1.89 (0.83)	1.66 (0.68)	1.75 (0.65)	1.65 (0.60)	2.98 (0.96)	2.22 (0.89)
No of Vars	21.00	22.34 (1.89)	22.49 (2.38)	22.22 (2.05)	22.55 (1.87)	25.26 (2.15)	28.00 (0.49)	98.23 (5.82)	24.78 (3.06)
No of Groups	4.00	4.33 (0.55)	4.37 (0.61)	4.49 (0.64)	4.43 (0.65)	4.20 (0.46)	4.00 (0.12)	5.48 (0.81)	5.39 (0.67)
Correct Group(%)	100.00	70.50	68.25	58.50	64.50	81.25	98.50	20.00	10.25
Correct Model(%)	100.00	23.50	28.50	18.00	26.00	1.50	0.00	0.00	3.00
FNR	0.00	1.35	1.19	2.33	0.81	0.21	0.07	0.00	1.88
FDR	0.00	6.74	6.94	7.12	7.11	16.50	25.04	43.51	15.80
					Example 2				
Model Error	0.00	1.39 (0.62)	1.36 (0.69)	1.56 (0.74)	1.26 (0.54)	1.61 (0.63)	1.66 (0.57)	2.96 (0.99)	1.94 (0.81)
No of Vars	18.00	19.29 (1.83)	19.46 (2.36)	19.30 (2.17)	19.11 (0.99)	22.65 (2.42)	28.03 (0.53)	38.21 (5.56)	21.60 (3.12)
No of Groups	4.00	4.30 (0.53)	4.34 (0.60)	4.45 (0.60)	4.35 (0.56)	4.13 (0.40)	4.00 (0.13)	5.48 (0.82)	5.24 (0.71)
Correct Group(%)	100.00	73.25	70.00	60.75	69.00	87.25	98.25	20.00	16.00
Correct Model(%)	100.00	27.75	35.75	23.50	45.25	1.50	0.00	0.00	4.00
FNR	0.00	1.09	0.75	2.11	0.33	0.21	0.05	0.03	2.18
FDR	0.00	7.04	7.15	7.85	5.58	19.84	35.80	51.56	17.03
					Example 3				
Model Error	0.00	1.69 (0.73)	1.63 (0.74)	1.86 (0.80)	1.67 (0.73)	1.75 (0.67)	1.64 (0.55)	3.01 (0.92)	2.22 (0.86)
No of Vars	21.00	22.52 (1.87)	22.18 (2.16)	22.12 (1.48)	22.59 (1.95)	25.32 (2.33)	28.03 (0.53)	38.63 (5.63)	24.69 (2.99)
No of Groups	4.00	4.45 (0.64)	4.34 (0.59)	4.54 (0.67)	4.64 (0.70)	4.95 (0.45)	5.00 (0.13)	5.80 (0.40)	5.43 (0.63)
Correct Group(%)	100.00	62.50	70.75	56.00	48.25	12.75	0.00	0.00	7.50
Correct Model(%)	100.00	20.75	32.75	19.75	23.50	1.50	0.00	0.00	2.75
FNR	0.00	1.04	1.23	2.29	0.86	0.19	0.05	0.00	1.82
FDR	0.00	7.19	5.79	6.62	7.28	16.58	25.09	44.21	15.48
					Example 4				
Model Error	0.00	0.38 (0.33)	0.33 (0.32)	0.33 (0.32)	0.25 (0.20)	0.46 (0.28)	0.77 (0.29)	1.90 (0.85)	0.31 (0.29)
No of Vars	4.00	5.18 (1.75)	4.90 (1.85)	4.98 (1.74)	4.40 (1.08)	6.51 (1.72)	16.00 (0.00)	17.26 (4.03)	4.75 (1.73)
No of Groups	2.00	2.31 (0.58)	2.24 (0.51)	2.34 (0.60)	2.18 (0.51)	2.01 (0.09)	2.00 (0.00)	2.16 (0.50)	2.40 (0.77)
Correct Group(%)	100.00	75.25	80.25	72.50	87.25	99.00	100.00	88.25	74.00
Correct Model(%)	100.00	50.25	63.25	57.25	78.75	10.25	0.00	0.00	65.50
FNR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR	0.00	16.57	12.16	13.72	6.12	34.33	75.00	76.10	10.51
					Example 5				
Model Error	0.00	1.07 (0.50)	1.04 (0.49)	1.13 (0.54)	0.91 (0.42)	0.87 (0.37)	0.80 (0.32)	1.68 (0.69)	1.5 (0.60)
No of Vars	16.00	16.99 (1.59)	16.71 (1.50)	16.88 (1.54)	16.46 (1.32)	16.47 (1.11)	16.00 (0.00)	16.46 (2.18)	17.24 (2.04)
No of Groups	2.00	2.67 (0.75)	2.48 (0.66)	2.65 (0.77)	2.34 (0.73)	2.27 (0.53)	2.00 (0.00)	2.06 (0.27)	2.36 (0.98)
Correct Group(%)	100.00	49.00	60.50	50.50	78.50	76.50	100.00	95.00	38.25
Correct Model(%)	100.00	41.00	51.25	42.00	74.75	76.50	100.00	95.00	27.00
FNR	0.00	0.63	0.78	0.97	0.23	0.00	0.00	0.00	1.36
FDR	0.00	5.74	4.42	5.51	2.55	2.46	0.00	1.78	7.49



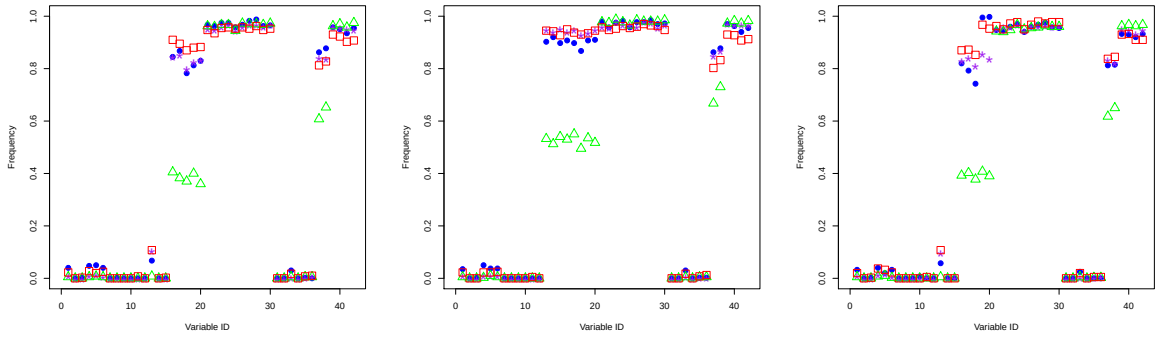
(a) Example 1 (b) Example 2 (c) Example 3

Figure 5.1: *Relative frequency plot of each covariate not been selected for sample size 400. Composite group bridge: solid blue dots; adaptive group bridge: filled red square; composite MCP: purple star; group bridge: green triangle point-up.*



(a) Example 1 (b) Example 2 (c) Example 3

Figure 5.2: *Relative frequency plot of each covariate not been selected for sample size 200. All the legends are same as in Figure 2.*



(a) Example 1 (b) Example 2 (c) Example 3

Figure 5.3: *Relative frequency plot of each covariate not been selected for sample size 100. All the legends are same as in Figure 2.*

Appendix A

Proofs

A.1 Proof of Propostion 3.2.1

Proof. For any fixed $\boldsymbol{\beta}$ and $\boldsymbol{\theta} \geq \mathbf{0}$,

$$\begin{aligned}\hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) &\equiv \arg\min_{\boldsymbol{\delta} \geq \mathbf{0}} \{S_{ln}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})\}, \\ &= \arg\min_{\boldsymbol{\delta} \geq \mathbf{0}} \left\{ \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \Big(\sum_{k \in A_j} \delta_k^{1-1/\mu} |\beta_k| + \psi \sum_{k \in A_j} \delta_k \Big) \right\}.\end{aligned}$$

Since the problem is separable in each δ_k, it follows that

$$\hat{\delta}_k = \hat{\delta}_k(\boldsymbol{\beta}, \boldsymbol{\theta}) = \arg\min_{\delta_k \geq \mathbf{0}} \left\{ \sum_{j: k \in A_j} \theta_j^{1-1/\gamma} c_j^{1/\gamma} (\delta_k^{1-1/\mu} |\beta_k| + \psi \delta_k) \right\} = (\frac{1-\mu}{\mu})^{\mu} \psi^{-\mu} |\beta_k|^{\mu}.$$

Because we have chosen $\psi = \mu^{\mu/(1-\mu)}(1-\mu)$,

$$\hat{\delta}_k = \mu^{\frac{\mu}{\mu-1}} |\beta_k|^{\mu}, \hat{\delta}_k^{1-1/\mu} = \mu |\beta_k|^{\mu-1}.$$

Substituting the above expressions to $S_{ln}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$, we see that

$$S_{ln}\{\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \Big(\sum_{k \in A_j} |\beta_k|^{\mu} \Big) + \tau_n \sum_{j=1}^{J} \theta_j. \qquad \text{(A.1.1)}$$

Therefore, to minimize $S_{ln}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta})$, it suffices to minimize $S_{ln}\{\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})\}$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta})$. For any fixed $\boldsymbol{\beta}$,

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{\beta}) &\equiv \arg\min_{\boldsymbol{\theta}\geq \mathbf{0}} \left[S_{ln}\{\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})\}\right] \\ &= \arg\min_{\boldsymbol{\theta}\geq \mathbf{0}} \left\{ \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} (\sum_{k\in A_j} |\beta_k|^{\mu}) + \tau_n \sum_{j=1}^{J} \theta_j \right\}.\end{aligned}$$

The problem is separable in each θ_j, and it follows that

$$\hat{\theta}_j = \hat{\theta}_j(\boldsymbol{\beta}) = \arg\min_{\theta_j\geq 0} \left(\theta_j^{1-1/\gamma} c_j^{1/\gamma} (\sum_{k\in A_j} |\beta_k|^{\mu}) + \tau_n \theta_j \right) = \tau_n^{-\gamma} (\frac{1-\gamma}{\gamma})^{\gamma} c_j (\sum_{k\in A_j} |\beta_k|^{\mu})^{\gamma}.$$

Substituting the above expression and $\tau_n = \lambda_n^{1/(1-\gamma)} \gamma^{\gamma/(1-\gamma)} (1-\gamma)$ to $S_{ln}\{\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})\}$, we see that

$$S_{ln}(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}), \hat{\boldsymbol{\delta}}(\boldsymbol{\beta}, \boldsymbol{\theta})) = \|\mathbf{y} - \sum_{k=1}^{d} \mathbf{x}_k \beta_k\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j (\sum_{k\in A_j} |\beta_k|^{\mu})^{\gamma},$$

which is exactly the same as the composite group bridge criterion. This completes the proof. □

A.2 Proof of Theorem 4.0.1

Proof. By the definition of $\hat{\boldsymbol{\beta}}_n$,

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2 \leq \lambda_n (\sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu\gamma} - \sum_{j=1}^{J} c_j \|\hat{\boldsymbol{\beta}}_{nA_j}\|_{\mu}^{\mu\gamma}).$$

Using $b^{\gamma} - a^{\gamma} \leq 2(b-a)b^{\gamma-1}$ for $0 \leq a \leq b$ and the Cauchy Schwartz inequality,

$$\begin{aligned}\|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu\gamma} - \|\hat{\boldsymbol{\beta}}_{nA_j}\|_{\mu}^{\mu\gamma} &\leq \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu\gamma} - \|\hat{\boldsymbol{\beta}}_{nA_j^1}\|_{\mu}^{\mu\gamma} \\ &\leq 2(\|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu} - \|\hat{\boldsymbol{\beta}}_{nA_j^1}\|_{\mu}^{\mu}) \cdot \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu(\gamma-1)} \\ &= 2\{\sum_{k\in A_j^1} (|\beta_{0k}|^{\mu} - |\hat{\beta}_{nk}|^{\mu})\} \cdot \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu(\gamma-1)} \\ &\leq 2\{2\sum_{k\in A_j^1} (|\beta_{0k} - \hat{\beta}_{nk}| \cdot |\beta_{0k}|^{\mu-1})\} \cdot \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu(\gamma-1)} \\ &\leq 4\|\hat{\boldsymbol{\beta}}_{nA_j^1} - \boldsymbol{\beta}_{0A_j^1}\|_2 \cdot \|\boldsymbol{\beta}_{0A_j^1}\|_{2\mu-2}^{\mu-1} \cdot \|\boldsymbol{\beta}_{0A_j^1}\|_{\mu}^{\mu(\gamma-1)}.\end{aligned}$$

Then we have,

$$\begin{aligned}\sum_{j=1}^{J} c_j\|\boldsymbol{\beta}_{0A_j}\|_\mu^{\mu\gamma} - \sum_{j=1}^{J} c_j\|\ \hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma} \leq& \sum_{j=1}^{J_1} c_j\|\boldsymbol{\beta}_{0A_j^1}\|_\mu^{\mu\gamma} - \sum_{j=1}^{J_1} c_j\|\ \hat{\boldsymbol{\beta}}_{nA_j^1}\|_\mu^{\mu\gamma}\\ \leq& 4\sum_{j=1}^{J_1} c_j\|\boldsymbol{\beta}_{0A_j^1}\|_{2\mu-2}^{\mu-1}\cdot\|\boldsymbol{\beta}_{0A_j^1}\|_\mu^{\mu(\gamma-1)}\cdot\|\hat{\boldsymbol{\beta}}_{nA_j^1}-\boldsymbol{\beta}_{0A_j^1}\|_2\\ \leq& 4\eta_n(\sum_{j=1}^{J_1}\|\hat{\boldsymbol{\beta}}_{nA_j^1}-\boldsymbol{\beta}_{0A_j^1}\|_2^2)^{1/2},\end{aligned}$$

where $\eta_n = (\sum_{j=1}^{J_1} c_j^2\|\boldsymbol{\beta}_{0A_j^1}\|_{2\mu-2}^{2\mu-2}\cdot\|\boldsymbol{\beta}_{0A_j^1}\|_\mu^{2\mu(\gamma-1)})^{1/2}$. Therefore,

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 - \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0\|_2^2 \leq& 4\lambda_n\eta_n(\sum_{j=1}^{J_1}\|\hat{\boldsymbol{\beta}}_{nA_j^1}-\boldsymbol{\beta}_{0A_j^1}\|_2^2)^{1/2}\\ \leq& 4\lambda_n\eta_n(\sum_{j=1}^{J}\|\hat{\boldsymbol{\beta}}_{nA_j}-\boldsymbol{\beta}_{0A_j}\|_2^2)^{1/2}\\ \leq& 4\lambda_n\eta_n\sqrt{c_n^*}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2.\\ 4\lambda_n\eta_n\sqrt{c_n^*}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2 \geq& \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 - \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0\|_2^2\\ =& \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\|_2^2 + 2\boldsymbol{\varepsilon}'\mathbf{X}(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n).\end{aligned}$$

The rest of the proof is very similar to Huang *et al.* (2009), and we shall present the details for the sake of completeness. Let $\delta_n = \|\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\|_2$ and $\boldsymbol{\varepsilon}_*$ be the projection of $\boldsymbol{\varepsilon}$ to the span of $\{\mathbf{X}_1,\ldots,\mathbf{X}_d\}$. By the Cauchy Schwarz inequality,

$$2|\boldsymbol{\varepsilon}'\mathbf{X}(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)| \leq 2\|\boldsymbol{\varepsilon}_*\|_2\delta_n \leq 2\|\boldsymbol{\varepsilon}_*\|_2^2 + \delta_n^2/2.$$

It follows that

$$\delta_n^2 - 4\lambda_n\eta_n\sqrt{c_n^*}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2 \leq 2|\boldsymbol{\varepsilon}'\mathbf{X}(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_n)| \leq 2\|\boldsymbol{\varepsilon}_*\|_2^2 + \delta_n^2/2,$$

and

$$\delta_n^2 \leq 4\|\boldsymbol{\varepsilon}_*\|^2 + 8\lambda_n\eta_n\sqrt{c_n^*}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2.$$

Since ρ_n is the smallest eigenvalue of $\mathbf{X}'\mathbf{X}/n$, the above inequality implies,

$$n\rho_n\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2^2 \leq \delta_n^2 \leq 4\|\boldsymbol{\varepsilon}_*\|_2^2 + 8\lambda_n\eta_n\sqrt{c_n^*}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2.$$

As $\boldsymbol{\varepsilon}_*$ is the projection of $\boldsymbol{\varepsilon}$ to a d-dimensional space, $\mathrm{E}\|\boldsymbol{\varepsilon}_*\|_2^2 \leq \sigma^2 d$. Thus,

$$\begin{aligned} E(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2) &\leq 4\sigma^2 d/(n\rho_n) + \frac{1}{2}\{8\lambda_n\eta_n\sqrt{c_n^*}/(n\rho_n)\}^2 + \frac{1}{2}E\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \\ &\leq \frac{8\sigma^2 d}{n\rho_n} + \frac{64\lambda_n^2\eta_n^2 c_n^*}{n^2\rho_n^2} \\ &= \frac{\sigma^2 d}{n\rho_n}(8 + 64c_n^* M_n), \end{aligned}$$

where $M_n = \frac{\lambda_n^2\eta_n^2}{n\rho_n\sigma^2 d} = O(1)$. This completes the proof. □

A.3 Proof of Theorem 4.0.2

Proof. Using the objective function and by the Karush-Kuhn-Tucker condition, for each $\hat{\beta}_{nk} \neq 0$,

$$2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n)'\mathbf{x}_k = \lambda_n\gamma\mu \sum_{j:k\in A_j} c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu(\gamma-1)} \cdot |\hat{\beta}_{nk}|^{\mu-1}\mathrm{sgn}(\hat{\beta}_{nk}).$$

Recall that $B_2 = \cup_{j=j_1+1}^{J} A_j$, $B_1 = B_2^c$. We define an oracle estimator,

$$\tilde{\boldsymbol{\beta}}_{nk} = \begin{cases} \hat{\boldsymbol{\beta}}_{nk} & k \notin B_2; \\ 0 & k \in B_2. \end{cases}$$

Then $(\hat{\beta}_{nk} - \tilde{\beta}_{nk})\mathrm{sgn}(\hat{\beta}_{nk}) = |\hat{\beta}_{nk}|I(k \in B_2)$. Note that even for $j \leq J_1$, $\hat{\boldsymbol{\beta}}_{nA_j} \neq \tilde{\boldsymbol{\beta}}_{nA_j}$, because of overlapping of predictors. It follows that

$$\begin{aligned} 2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n)'\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n) &= \lambda_n\gamma\mu \sum_{k\in B_2} |\hat{\beta}_{nk}| \cdot |\hat{\beta}_{nk}|^{\mu-1} \sum_{j:k\in A_j} c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu(\gamma-1)} \\ &= \lambda_n\gamma\mu \sum_{j=1}^{J} c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu(\gamma-1)}(\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^\mu - \|\tilde{\boldsymbol{\beta}}_{nA_j}\|_\mu^\mu). \end{aligned}$$

Since $\gamma b^{\gamma-1}(b-a) \leq b^\gamma - a^\gamma$ for $0 \leq a \leq b$, for $j \leq J_1$, we have

$$\gamma\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu(\gamma-1)}(\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^\mu - \|\tilde{\boldsymbol{\beta}}_{nA_j}\|_\mu^\mu) \leq \|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma} - \|\tilde{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}.$$

It follows that,

$$|2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n)'\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}_n)| \leq \lambda_n\mu \sum_{j=1}^{J_1} c_j(\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma} - \|\tilde{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}) + \lambda_n\gamma\mu \sum_{j=J_1+1}^{J} c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}.$$

Therefore,

$$
\begin{aligned}
&|(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_n)'\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\tilde{\boldsymbol{\beta}}_n)|+(1-\gamma)\lambda_n\mu\sum_{j=J_1+1}^{J}c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}\\
\leq&\lambda_n\mu\sum_{j=1}^{J}c_j(\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}-\|\tilde{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma})\\
\leq&\mu\|\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}_n\|_2^2-\mu\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2\\
=&\mu\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\tilde{\boldsymbol{\beta}}_n)\|_2^2+2\mu(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_n)'\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\tilde{\boldsymbol{\beta}}_n)
\end{aligned}
$$

Since $n\rho_n^*$ is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$, it follows that

$$
(1-\gamma)\lambda_n\sum_{j=J_1+1}^{J}c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}\leq\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n-\tilde{\boldsymbol{\beta}}_n)\|_2^2\leq n\rho_n^*\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2^2\leq n\rho_n^*\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2^2.
$$

We also have

$$
\begin{aligned}
\sum_{j=J_1+1}^{J}c_j\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}&\geq\sum_{j=J_1+1}^{J}\|\hat{\boldsymbol{\beta}}_{nA_j}\|_\mu^{\mu\gamma}\\
&\geq(\sum_{j=J_1+1}^{J}\sum_{k\in A_j}|\hat{\beta}_{nk}|^\mu)^\gamma\\
&\geq\{(\sum_{k\in B_2}|\hat{\beta}_{nk}|^\mu)^{1/\mu}\}^{\mu\gamma}\\
&\geq\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2^{\mu\gamma}.
\end{aligned}
$$

Altogether, we have

$$
(1-\gamma)\lambda_n\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2^{\mu\gamma}\leq n\rho_n^*\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2^2\leq n\rho_n^*\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2^2\leq O_p(\sigma^2d\rho_n^*/\rho_n).
$$

This implies that when $\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2>0$,

$$
(1-\gamma)\lambda_n\leq n\rho_n^*(\frac{\sigma^2d}{n\rho_n})^{1-\mu\gamma/2}.
$$

Therefore,

$$
\begin{aligned}
P(\|\hat{\boldsymbol{\beta}}_{nB_2}\|_2>0)&\leq P\left\{(1-\gamma)\lambda_n\leq n\rho_n^*(\frac{\sigma^2d}{n\rho_n})^{1-\mu\gamma/2}O_p(1)\right\}\\
&=P\left\{\frac{\lambda_n\rho_n^{1-\mu\gamma/2}}{d^{1-\mu\gamma/2}\rho_n^*n^{\mu\gamma/2}}\leq O_p(1)\right\}\to 0,
\end{aligned}
$$

as $n\to\infty$, based on assumption A3. This completes the proof. □

A.4 Proof of Theorem 4.0.3

Proof. Based on the estimation error bound, we know that for sufficiently large C, $\hat{\boldsymbol{\beta}}_n$ lies in the ball $\{\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq h_n C\}$ with probability 1, where $h_n = \{d/(n\rho_n)\}^{1/2}$. Define $\boldsymbol{\beta}_{1n} = \boldsymbol{\beta}_{0B_1^1} + h_n \mathbf{u}_1$ and $\boldsymbol{\beta}_{2n} = \boldsymbol{\beta}_{0B_1^2} + h_n \mathbf{u}_2$ with $\|\mathbf{u}_1\|_2^2 + \mathbf{u}_2\|_2^2 = \|\mathbf{u}\|_2^2 \leq C^2$. Define

$$\begin{aligned} V_n(\mathbf{u}_1, \mathbf{u}_2) &= L_n\{(\boldsymbol{\beta}_{1n}', \boldsymbol{\beta}_{2n}', \mathbf{0})'\} - L_n\{(\boldsymbol{\beta}_{0B_1^1}', \mathbf{0}', \mathbf{0}')'\} \\ &= L_n\{(\boldsymbol{\beta}_{0B_1^1}' + h_n \mathbf{u}_1', h_n \mathbf{u}_2', \mathbf{0}')'\} - L_n\{(\boldsymbol{\beta}_{0B_1^1}', \mathbf{0}', \mathbf{0}')'\}. \end{aligned}$$

By the group selection consistency results established in Theorem 2, it holds with probability tending to 1 that

$$\arg \min_{\mathbf{u}: \|\mathbf{u}\|_2 \leq C} V_n(\mathbf{u}_1, \mathbf{u}_2) = \{h_n^{-1}(\hat{\boldsymbol{\beta}}_{nB_1^1} - \boldsymbol{\beta}_{0B_1^1}), h_n^{-1}(\hat{\boldsymbol{\beta}}_{nB_1^2} - \boldsymbol{\beta}_{0B_1^2})\}.$$

That is, we only have to consider the predictors in B_1. Now to establish the desired result, it suffices to show that if $\|\mathbf{u}_2\|_2 > 0$ for any $\|\mathbf{u}\|_2 \leq C$, $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) > 0$ with probability tending to 1.

Recall that $\mathbf{X}_{11} = (\mathbf{x}_k, k \in B_1^1)$ and $\mathbf{X}_{12} = (\mathbf{x}_k, k \in B_1^2)$. It can be shown that

$$\begin{aligned} & V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) \\ =& h_n^2 \mathbf{u}_2' \mathbf{X}_{12}' \mathbf{X}_{12} \mathbf{u}_2 - 2h_n^2 \mathbf{u}_2' \mathbf{X}_{12}' \mathbf{X}_{11} \mathbf{u}_1 + 2h_n \mathbf{u}_2' \mathbf{X}_{12}' \varepsilon \\ & + \lambda_n \sum_{j=1}^{J_1} c_j \left\{ (\sum_{k \in A_j^1} |\beta_{0k} + h_n u_{1k}|^\mu + \sum_{k \in A_j^2 \cap B_1} |h_n u_{2k}|^\mu)^\gamma - (\sum_{k \in A_j^1} |\beta_{0k} + h_n u_{1k}|^\mu)^\gamma \right\} \\ =& T_{1n} + T_{2n} + T_{3n} + T_{4n}. \end{aligned}$$

For the first two terms, we have

$$\begin{aligned} T_{1n} + T_{2n} &\geq h_n^2 \mathbf{u}_2' \mathbf{X}_{12}' \mathbf{X}_{12} \mathbf{u}_2 - h_n^2 (\mathbf{u}_2' \mathbf{X}_{12}' \mathbf{X}_{12} \mathbf{u}_2 + \mathbf{u}_1' \mathbf{X}_{11}' \mathbf{X}_{11} \mathbf{u}_1) \\ &= -h_n^2 \mathbf{u}_1' \mathbf{X}_{11}' \mathbf{X}_{11} \mathbf{u}_1 \\ &\geq -n h_n^2 \tau_{1n}^* \|\mathbf{u}_1\|_2^2 \\ &\geq -\tau_1^* (d/\rho_n) C^2 \end{aligned}$$

For the third term, because

$$\begin{aligned} E(\mathbf{u}_2'\mathbf{X}_{12}'\varepsilon) &\le \{E(\mathbf{u}_2'\mathbf{X}_{12}'\boldsymbol{\varepsilon})^2\}^{1/2} \\ &= \{\sigma^2 \mathrm{tr}(\mathbf{X}_{12}\mathbf{u}_2\mathbf{u}_2'\mathbf{X}_{12}')\}^{1/2} \\ &= \sigma(\mathbf{u}_2'\mathbf{X}_{12}'\mathbf{X}_{12}\mathbf{u}_2)^{1/2} \\ &\le \sigma n^{1/2}\rho_n^{*1/2}\|\mathbf{u}_2\|_2, \end{aligned}$$

then we have

$$T_{3n} = h_n n^{1/2}\rho_n^{*1/2}\|\mathbf{u}_2\|O_p(1) = d^{1/2}(\rho_n^*/\rho_n)^{1/2}O_p(1).$$

Now consider the forth term. Because $b^\gamma - a^\gamma \ge \gamma b^{\gamma-1}(b-a)$ for $0 \le a \le b$, we have

$$\begin{aligned} &\lambda_n \sum_{j=1}^{J_1} c_j \left\{ (\sum_{k\in A_j^1} |\beta_{0k} + h_n u_{1k}|^\mu + \sum_{k\in A_j^2\cap B_1} |h_n u_{2k}|^\mu)^\gamma - (\sum_{k\in A_j^1} |\beta_{0k} + h_n u_{1k}|^\mu)^\gamma \right\} \\ \ge &\lambda_n \gamma \sum_{j=1}^{J_1} c_j \left(\sum_{k\in A_j^2\cap B_1} |h_n u_{2k}|^\mu \right) \left(\sum_{k\in A_j^1} |\beta_{0k} + h_n u_{1k}|^\mu + \sum_{k\in A_j^2\cap B_1} |h_n u_{2k}|^\mu \right)^{\gamma-1} \\ = &\lambda_n \gamma h_n^\mu O(\|\mathbf{u}_2\|_\mu^\mu)\cdot O(1) \\ \ge &\lambda_n h_n^\mu O(\|\mathbf{u}_2\|_2^\mu) \\ = &\frac{\lambda_n \rho_n^{1-\mu/2}}{d^{1-\mu/2}\rho_n^* n^{\mu/2}} \cdot \frac{d\rho_n^*}{\rho_n} \cdot O(\|\mathbf{u}_2\|_2^\mu). \end{aligned}$$

By A3*, if $\|\mathbf{u}_2\|_2 \neq 0$, the fourth term T_{4n} dominates the first three terms and $T_{4n} \to \infty$ as $n \to \infty$; consequently, $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}) > 0$ with probability tending to 1. This completes the proof. □

A.5 Proof of Theorem 4.0.4

Proof. Since $[B_1^1, \boldsymbol{\beta}_{0B_1^1}, J_1]$ are fixed, (4.0.2) implies that assumptions A2 and A3 hold. Recall that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0B_1^1}', \boldsymbol{\beta}_{0B_1^2}', \boldsymbol{\beta}_{0B_2}')'$ and $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{nB_1^1}', \hat{\boldsymbol{\beta}}_{nB_1^2}', \hat{\boldsymbol{\beta}}_{nB_2}')'$. Since all the assumptions of Theorems 4.0.1 and 4.0.2 are met, we have $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 = O_p(1/n)$ and $\|\hat{\boldsymbol{\beta}}_{nB_1} - \boldsymbol{\beta}_{0B_1}\|_2^2 =$

$O_p(1/n)$. Let $h_n{=}n^{-1/2}$. For $\mathbf{u}_1 \in \mathbb{R}^{|B_1^1|}$, $\mathbf{u}_2 \in \mathbb{R}^{|B_1^2|}$ and $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')'$, define

$$\begin{aligned} V_{1n}(\mathbf{u}) &= L_n(\boldsymbol{\beta}_0 + h_n(\mathbf{u}_1', \mathbf{u}_2', \mathbf{0}_{B_2}')') - L_n(\boldsymbol{\beta}_0) \\ &= \left\{-2h_n\mathbf{u}'\mathbf{X}_1'\varepsilon + h_n^2\mathbf{u}'\mathbf{X}_1'\mathbf{X}_1\mathbf{u}\right\} + \lambda_n \sum_{j=1}^{J_1} c_j \left\{ \Big(\sum_{k \in A_j \cap B_1} |\beta_{0k} + h_n u_k|^{\mu} \Big)^{\gamma} - \|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu\gamma} \right\} \\ &= T_{1n}(\mathbf{u}) + T_{2n}(\mathbf{u}). \end{aligned}$$

Let $\hat{\mathbf{u}}_n = \arg\min\{V_{1n}(\mathbf{u}), \mathbf{u} \in \mathbb{R}^{|B_1|}\}$. By the group selection consistency results established in Theorem 4.0.2, we have $\sqrt{n}\hat{\boldsymbol{\beta}}_{nB_2} \to_d \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = h_n(\hat{\mathbf{u}}_{n1}', \hat{\mathbf{u}}_{n2}', \mathbf{0}_{B_2}')'$ with probability tending to 1.

Consider the first term $T_{1n}(\mathbf{u})$,

$$T_{1n}(\mathbf{u}) \to_d -2\mathbf{u}'\mathbf{W}_1 + \mathbf{u}'\boldsymbol{\Sigma}_1\mathbf{u}.$$

Now consider the second term $T_{2n}(\mathbf{u})$. For each $j = 1, ..., J_1$,

$$\begin{aligned} &\lim_{h_n \to 0} \left\{ \lambda_n h_n \frac{(\sum_{k \in A_j \cap B_1} |\beta_{0k} + h_n u_k|^{\mu})^{\gamma} - \|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu\gamma}}{h_n} \right\} \\ =& \lambda_0\gamma\|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu(\gamma-1)} \sum_{k \in A_j \cap B_1} \lim_{h_n \to 0} \left\{ \frac{|\beta_{0k} + h_n u_k|^{\mu} - |\beta_{0k}|^{\mu}}{h_n} \right\} \\ =& \begin{cases} \lambda_0\gamma\|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu(\gamma-1)} \sum_{k \in A_j^1} u_k \mu |\beta_{0k}|^{\mu-1}\mathrm{sgn}(\beta_{0k}) & u_k = 0 \text{ for any } k \in A_j^2 \cap B_1; \\ \infty & \text{otherwise.} \end{cases} \end{aligned}$$

It follows that

$$T_{2n}(\mathbf{u}) \to \begin{cases} \lambda_0\gamma\mu \sum_{j=1}^{J_1} c_j \|\boldsymbol{\beta}_{0A_j}\|_{\mu}^{\mu(\gamma-1)} \sum_{k \in A_j^1} u_k |\beta_{0k}|^{\mu-1}\mathrm{sgn}(\beta_{0k}) & u_k = 0 \text{ for any } k \in B_1^2; \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we must have $\sqrt{n}\hat{\boldsymbol{\beta}}_{nB_1^2} \to_d \mathbf{0}$; we note that a strong condition in A3* is not needed to establish this result. The rest of the results in Theorem 4.0.3 follows from the limiting objective function and by invoking the argmax theorem in, e.g., van der Vaart (2000). This completes the proof. □

Appendix B

R code

B.1 R code for Composite group bridge estimation

```
library(MASS)    ## Load library MASS
library(lars)     ## Load library LARS

#####################################################################
Creating a Function named BridgeBridge with Y:response
vector,X:Design matrix,G:Group matrix lambda value,gamma
value,mu value and initial value of beta as input parameters

#####################################################################

BridgeBridge <- function(Y,X,G,lambda,ini=NULL,gamma=0.5,mu=0.5)
{

    n <- nrow(X) ## Assign n as the number of observations
    p <- ncol(X) ## Assign p as the number of predictor variables
```

```
J <- ncol(G) ## J :No of groups

# Least square estimate as the initial estimate
beta_ls <- ginv(t(X)%*%X)%*%t(X)%*%Y

# Calculate group level weights Step 1
theta1=(lambda*gamma)^(gamma/(gamma-1))

# Calculate tau
tau <- lambda^{1/(1-gamma)}*gamma^{gamma/(1-gamma)}*(1-gamma)
h <- apply(G,1,sum)

# if the initial value is null or the estimated beta has all zeros
# then assign beta_p as Least square estimate else assign the
# beta_p as the initial value
 if(is.null(ini) | sum(ini)==0)
     {
        beta_p <- beta_ls
     }
else
     {
        beta_p <- ini
     }
beta_p <- as.matrix(beta_p,nrow=p)
diff <- 1
j <- 1
zero <- FALSE
```

```
while(diff > 1e-6)
    {
    ## Calculate group level weights(unpowered) Step 2:betaJ
     betaJ <- apply(G*as.vector(abs(beta_p)^mu),2,sum)
     c <- apply(G,2,sum)^(1-gamma)

    ## Calculating group level weights(un powered)
    ## Step 3 :theta1*c*betaJ^gamma
    theta <- theta1*c*betaJ^gamma

    ## Calculating individual level weights(unpowered)
    delta=mu^(mu/(mu-1))*abs(beta_p)^mu/h

    ## if theta=0, then the corresponding predictors shall be removed.
    ##  gid :ids of zero groups
    gid <- which(theta==0)
    if(length(gid)==0)
        {
         gid <- J+1
        }

    ## pid:ids of zero predictors
    pid <- which(delta==0)
    if(length(pid)==0)
        {
        pid <- p+1
```

```
    }

## Group level weights calculation (significant groups)
wJ <- theta[-gid]^(1-1/gamma)*c[-gid]^(1/gamma)

## Individual level weights calculation( significant predictors )
wI=delta[-pid]^(1-(1/mu))

## Calculating wlasso weights
wlasso
     <- apply(G[-pid,-gid]%*%diag(wJ,nrow=length(wJ))*as.vector(wI),1,sum)

## fitting the model using lars package
## with  Intercept option:False,normalize:False
## transforming the wlasso by inverting the weights
fit
     <- lars(X[,-pid]%*%diag(wlasso^(-1),nrow=length(wlasso)),
          Y,intercept=FALSE,normalize=FALSE)

beta_c
     <- as.matrix(predict.lars(fit,s=lambda/(2/n),
          type="coe",mode="lambda")$coe)

## transforming back beta_c : dividing by wlasso
beta_c <- beta_c/wlasso

## Calculating the difference between current beta
```

```
    ## and the beta calculated in the previous iteration
    diff <- sum((beta_p[-pid]-beta_c)^2)/sum(beta_p[-pid]^2)
    beta_p <- rep(0,p)
    beta_p[-pid] <- beta_c
    j <- j+1

    ## if all the predictors are zero come out of the while loop
      if(sum(beta_p==0)==p)
        {
          diff <- 1e-6
          zero <- TRUE
        }
} ## close while loop

## If all the predictors are zero
 if(zero==TRUE)
    {
    ## Show that degrees of freedom are =0
    df1 <- 0
    df2 <- 0
    gcount <- 0                 ## Number of significant groups=0
    groupids <- NA              ## No significant group ids
    sigvarid <- NA              ## No significant variable ids
    sigvaridcount <- 0          ##  Number of significant variables =0
    }
else
   {
```

```
    ##calculating the degrees of freedom
    betaJ <- apply(G*as.vector(abs(beta_p))^mu,2,sum)
    c <- apply(G,2,sum)^(1-gamma)
    theta <- theta1*c*betaJ^gamma
    wJ <- theta^(1-1/gamma)*c^{1/gamma}
    gid <- which(theta==0)
    gcount<-length(which(theta!=0))
    groupids<-which(theta!=0)
    if(length(gid)==0) gid <- J+1
    pid <- which(beta_p==0)
    sigvaridcount<-length(which(beta_p!=0))   ## count of significant predictors
    sigvarid<-which(beta_p!=0)                ## ids of significant predictors
    if(length(pid)==0) pid <- p+1
    Wlam <- apply(as.matrix(G[-pid,-gid]*as.vector(abs(beta_p[-pid]))),1,sum)
    Xlam <- X[,-pid]

    ## calculating degrees of freedom1
    df1 <- sum(diag(Xlam%*%ginv(t(Xlam)%*%Xlam
                    +0.5*diag(Wlam,nrow=length(Wlam)))%*%t(Xlam)))

    ## degrees of freedom 2 is the number of active non zero variables
    df2 <- sum(beta_p!=0)

    }
logsse <- log(sum((Y-X%*%beta_p)^2)/n)

## calculating BIC penalty 1 using degrees of freedom1
```

```
     bic1 <- logsse + log(n)/n*df1

    ## calculating BIC2 using degrees of freedom 2
    bic2 <- logsse + log(n)/n*df2

    ## calculating AIC penalty 1 using degrees of freedom1
    aic1 <- logsse + 2/n*df1

    ## calculating AIC2 using degrees of freedom 2
    aic2 <- logsse + 2/n*df2

    list(bic=bic1,aic=aic1,bic2=bic2,aic2=aic2,df=c(df1,df2),
                  beta=beta_p,noofgroups=gcount,siggroups=groupids,
                  sigvar=sigvarid,sigvarcount=sigvaridcount)
}
```

B.2 R code for Adaptive Group Bridge estimation

```
library(MASS)
AdaGB <- function(Y,X,G,lambda,ini=NULL,gamma=0.5,mu=2,ada=TRUE)
{

    n <- nrow(X)
    p <- ncol(X)
    J <- ncol(G)
    beta_ls <- ginv(t(X)%*%X)%*%t(X)%*%Y
    if(ada==TRUE)
```

```
{
    ##predictor occurences
    h <- apply(G,1,sum)

    ##weights
    w <- abs(beta_ls)^{-mu}/h

    ##c_j
    c <- apply(G*as.vector(abs(beta_ls*w)),2,sum)^(1-gamma)   ###c_1
    ##c <- apply(G,2,sum)^(1-gamma)                           ###c_2
}
else
{
    w <- rep(1,p)
    c <- apply(G,2,sum)^(1-gamma)
}

##tau
tau <- lambda^{1/(1-gamma)}*gamma^{gamma/(1-gamma)}*(1-gamma)
theta_1 <- ifelse(gamma==1,1,((1-gamma)/gamma/tau)^gamma*c)
if(is.null(ini))
{
    beta_p <- beta_ls
}
else
{
    beta_p <- ini
```

```
}
beta_p <- as.matrix(beta_p,nrow=p)
diff <- 1
j <- 1
zero <- FALSE
while(diff > 1e-6)
    {
    betaJ <- apply(G*as.vector(abs(beta_p*w)),2,sum)
    theta <- theta_1*betaJ^gamma

    ##if theta=0, then the corresponding predictors shall be removed.
    gid <- which(theta==0)
    if(length(gid)==0)
    {
        gid <- J+1
        pid <- p+1
    }
    else
    {
      pid <- which(apply(as.matrix(G[,gid]),1,sum)!=0)
    }

    ##calculate weights for adLasso
    if(gamma==1)
    {
        wJ <- lambda*c[-gid]^(1/gamma)
    }
```

```
    else
    {
        wJ <- theta[-gid]^(1-1/gamma)*c[-gid]^(1/gamma)
    }

    ##if gamma=1, the same as w.
    wlasso <- apply(G[-pid,-gid]%*%
                    diag(wJ,nrow=length(wJ))*as.vector(w[-pid]),1,sum)
    fit <- lars(X[,-pid]%*%diag(wlasso^(-1),
             nrow=length(wlasso)),Y,intercept=FALSE,normalize=FALSE)
    beta_c <- as.matrix(predict.lars
                   (fit,s=lambda/2/n,type="coe",mode="lambda")$coe)
    beta_c <- beta_c/wlasso
    diff <- sum((beta_p[-pid]-beta_c)^2)/sum(beta_p[-pid]^2)
    beta_p <- rep(0,p)
    beta_p[-pid] <- beta_c
    j <- j+1
    if(sum(beta_p==0)==p)
    {
        diff <- 1e-6
        zero <- TRUE
    }
}

if(zero==TRUE)
{
    df1 <- 0
```

```
    df2 <- 0

}
else
{
    ##calculate df
    betaJ <- apply(G*as.vector(abs(beta_p*w)),2,sum)
    theta <- theta_1*betaJ^gamma
    if(gamma==1)
    {
        wJ <- lambda*c^(1/gamma)
    }
    else
    {
        wJ <- theta^(1-1/gamma)*c^(1/gamma)
    }

    gid <- which(theta==0)
    gcount<-length(which(theta!=0))
    groupids<-which(theta!=0)
    if(length(gid)==0) gid <- J+1
    pid <- which(beta_p==0)
    sigvaridcount<-length(which(beta_p!=0))
    sigvarid<-which(beta_p!=0)
    if(length(pid)==0) pid <- p+1
    Wlam
        <- apply(as.matrix(G[-pid,-gid]*as.vector(abs(beta_p[-pid]))),1,sum)
```

```
        Xlam <- X[,-pid]
        df1 <- sum(diag(Xlam%*%ginv(t(Xlam)%*%Xlam
                    +0.5*diag(Wlam,nrow=length(Wlam)))%*%t(Xlam)))
        df2 <- sum(beta_p!=0)
    }

    logsse <- log(sum((Y-X%*%beta_p)^2)/n)
    bic1 <- logsse + log(n)/n*df1
    bic2 <- logsse + log(n)/n*df2
    aic1 <- logsse + 2/n*df1
    aic2 <- logsse + 2/n*df2

    list(bic=bic1,aic=aic1,bic2=bic2,aic2=aic2,df=c(df1,df2),beta=beta_p,
                noofgroups=gcount,siggroups=groupids,sigvar=sigvarid,
                sigvarcount=sigvaridcount)
}
```

Bibliography

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Breheny, P. and Huang, J. (2009a) Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2**, 369–380.

Breheny, P. and Huang, J. (2009b) Penalized methods for bi-level variable selection. *Statistics and its interface*, **2**, 369–380.

Breheny, P. J. (2009) *Regularized methods for high-dimensional and bi-level variable selection*. Ph.D. thesis, University of Iowa.

Buhlmann, P. and van de Geer, S. (2009) *Statistics for High-Dimensional Data*. Springer.

Efron, B., Hastie, T., Johnstones, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, **32(2)**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Friedman, J., Trevor, H., Hfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. Tech. rep., Annals of Applied Statistics.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference and prediction*. Springer.

Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high dimensional models. URL http://arxiv.org/abs/1204.6491.

Huang, J., Horowitz, J. L. and Ma, S. (2008a) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, **36**, 587–613.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009) A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

Huang, J., Ma, S. and Zhang, C.-H. (2008b) Adaptive lasso for high-dimensional regression models. *Statistica Sinica*, **18**, 1603–1618.

Izenman, A. J. (2008) *Mordern Multivariate Statistical Techniques*. Springer.

Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators. *The Annals of Statistics*, **28**, 1356–1378.

Mallows, C. L. (1973) Some comments on cp. *Technometrics*, **15**, pp. 661–675. URL http://www.jstor.org/stable/1267380.

Meier, L., van de Geer, S. and Buhlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 53–71. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x.

Park, Y. M. and Hastie, T. (2007) L1-regularization path algorithm for generalized linear models. *Royal Statistical Society*, **69**, 659–677.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Stone, M. (1974) Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–515.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, **58**, 267–288.

van der Vaart, A. W. (2000) *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Wang, L., Chen, G. and Li, H. (2007) Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486–1494. URL http://bioinformatics.oxfordjournals.org/content/23/12/1486.abstract.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Zhang (2007) Penalized linear unbiased selection. Tech. rep.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567–1594.

Zhao, P., Rocha, G. and Yu, B. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist*, **37**, 3468–3497.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, **67**, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the degree of freedom of the lasso. *The Annals of Statistics*, **35**, 2173–2192.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *ANNALS OF STATISTICS*, **36**, 1509.

Zou, H. and Zhang, H. H. (2009) On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, **37**, 1733.